Available Online: 9 February 2026 DOI: 10.54254/2977-3903/2026.31770

Gated fusion and multi-task learning for multimodal emotion recognition in meme-rich contexts

Xufeng Liu

Queen Mary University of London, London, UK

catkinsflow@163.com

Abstract. Multimodal emotion recognition plays a vital role in social media analysis. However, traditional systems face a fundamental challenge: cross-modal semantic conflicts triggered by internet memes lead to a severe negative transfer effect—significantly degrading the model's ability to recognise emotions even in regular non-meme content. This study systematically reveals and quantifies this phenomenon, providing critical insights for overcoming the bottleneck of multimodal recognition in meme-rich contexts. To investigate this issue, we constructed a balanced mixed dataset and first empirically confirmed that conventional fusion methods suffer a substantial performance drop under this setting. In response, we propose the Meme-Adaptive Gated Fusion Network (MAG-Net), which is built upon shared Bidirectional Encoder Representations from Transformers (BERT)-Residual Neural Network (ResNet) encoders, incorporates a gated fusion module to dynamically adjust the fusion weights of image and text features, and employs a multi-task learning framework to jointly optimise emotion classification and meme detection. Through feature sharing and the adaptive modulation of fusion ratios via the gating network, our method effectively mitigates cross-modal conflicts. Our experiments demonstrate that on a test set containing 50% memes, the method achieves 56.93% emotion recognition accuracy, outperforming the baseline by 14.65%, while reaching 94.02% meme detection accuracy. Importantly, our framework not only improves overall performance but also restores the model's ability to correctly process regular content, thereby recovering the advantages of multimodal fusion in challenging meme-rich scenarios.

Keywords: memes, multimodal emotion recognition, gated fusion, multi-task learning

1. Introduction

In contemporary social media, memes have become a dominant medium of visual communication. Through their fixed format of "image template + textual caption" combined with rhetorical devices such as satire, exaggeration, and metaphor, memes achieve a highly condensed expression of information and emotion, playing a key role in shaping public opinion, brand communication, and subcultures [1-3]. While memes exhibit strong communicative power, they pose unprecedented challenges for automated sentiment analysis [4, 5].

The core challenge of memes stems from their unique logic of meaning construction [6]. First, there is often a direct conflict between image and text sentiment. Meme images are frequently remixed and adapted, so their surface emotion (e.g., a smiling face) often contradicts the true intent of the text (e.g., sarcastic criticism), resulting in clear cross-modal irony [7, 8]. Second, meme understanding depends heavily on context. Interpretation relies on the shared knowledge and implicit background of specific communities, making it difficult for multimodal models that rely only on surface semantics to capture their true emotional tendency.

Multimodal Emotion Recognition (MER) aims to integrate textual and visual information to infer emotional states more accurately, and has seen significant progress in recent years. Various fusion methods have emerged, including attention mechanisms, cross-modal alignment, and unified Transformer-based representations [9-12]. However, most existing methods are built on a strong implicit assumption: that the image and text modalities are semantically consistent and highly correlated in emotional expression. This assumption may hold for conventional social media content, but breaks down for memes, where the relationship between image and text is often intentionally disjointed or even contradictory. For example, in hateful meme detection tasks, researchers have observed that multimodal models struggled to effectively fuse contradictory image-text signals [13].

This study proposes that memes are not only difficult to classify, but also introduce a "negative transfer" effect that systematically harms a model's ability to recognise emotion in regular content. When a model is trained on a mixture of meme and non-meme posts, the widespread image-text conflict in meme samples interferes with the decision boundaries learned by the model, leading to a significant drop in performance even on unseen regular content. Although memes are pervasive on social media, this negative effect and its underlying mechanisms have not yet been systematically verified or quantified in multimodal emotion recognition.

To address these challenges, this paper proposes MAG-Net, a framework specifically tailored for meme-rich scenarios. Rather than pursuing a generic, state-of-the-art model architecture, MAG-Net is designed as a lightweight, adaptive, and efficient solution that explicitly targets the image-text conflict and negative transfer caused by memes.

Specifically, we formulate meme detection as an auxiliary task and jointly optimise it with the primary emotion recognition task under a shared feature encoding and fusion architecture. In addition, we introduce a gating network that enables the model to adaptively adjust the fusion weights of textual and visual features based on the input instance, thereby dynamically reducing the influence of the noisy modality when image-text conflict occurs.

The main contributions of this work are as follows:

1. Identifying a Critical Issue

We provide the first empirical evidence that internet memes induce a strong negative transfer effect in multimodal emotion recognition. Experimental results show that models trained on meme-mixed data suffer substantial performance degradation, even on non-meme content.

2. Proposing a Dedicated Framework

We propose a novel gated fusion multi-task learning framework to mitigate this issue. The framework combines an adaptive gating mechanism that dynamically balances textual and visual cues with an auxiliary meme detection task that helps the model learn conflict patterns.

3. Demonstrating Practical Effectiveness

Extensive experiments on a realistic mixed dataset demonstrate that our approach not only outperforms standard fusion methods but, crucially, restores the model's original performance on regular social media posts. This makes MAG-Net a robust and practical solution for emotion analysis in meme-rich online environments.

2. Related work

2.1. Multimodal emotion recognition

MER aims to integrate textual and visual information to infer a user's emotional state more accurately. Early studies primarily relied on feature concatenation or shallow fusion strategies, whereas recent work has widely adopted attention mechanisms, gated fusion units, and Transformer-based unified representations to enhance inter-modal interaction [14-16]. These approaches have achieved strong performance on standard image-text datasets; however, they are largely built on a key assumption—that the image and text are highly consistent and semantically aligned in emotional expression. While this assumption generally holds for conventional social media content, it breaks down in the presence of multimodal semantic conflicts, such as those commonly found in memes.

2.2. Meme analysis and multimodal challenges

Memes have become an important research object in computational social science and multimedia analysis due to their unique communicative properties and rhetorical structures, including irony, exaggeration, and metaphor. Existing research has mainly focused on three directions.

- Harmful Content Detection

Benchmark studies such as the Hateful Memes Challenge have demonstrated that semantic contradictions between images and text in memes can cause multimodal models to fail. These tasks focus on distinguishing hateful from non-hateful memes, but do not involve fine-grained sentiment or emotion analysis [13].

- Sarcasm Detection

Research on multimodal sarcasm detection has shown that image-text incongruity is a key cue for identifying sarcasm. However, such work focuses on a binary classification (sarcastic or not) rather than continuous sentiment dimensions or specific emotion categories [7, 8].

- Dataset Construction

Resources such as MET-Meme and MultiOFF provide annotated meme datasets covering aspects such as intent, metaphor type, or offensiveness. However, their emotion labels are mostly coarse-grained or absent, limiting their applicability to emotion recognition tasks [17, 18].

Research Gap: Although prior work highlights the unique challenges posed by memes in multimodal understanding, it does not systematically examine their impact on conventional emotion recognition tasks. In particular, existing studies neither quantify the potential negative transfer effect induced by memes nor propose dedicated solutions to mitigate image-text emotion conflict.

2.3. Multi-task learning and adaptive fusion

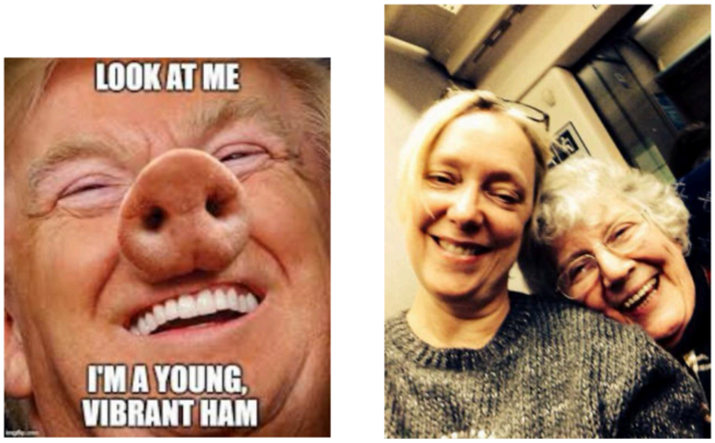
Multi-task learning improves model generalisation through shared representations [19]. It has often been used in multimodal settings to jointly learn emotion classification and related auxiliary tasks (e.g., emotion intensity prediction) [20]. However, existing approaches typically select auxiliary tasks that are semantically aligned with the main task and have not explored meme detection as an auxiliary task to improve the robustness of emotion recognition [21].

With respect to fusion strategies, adaptive gating mechanisms have been proposed to dynamically modulate multimodal information flow. These mechanisms were originally designed to handle missing modalities or general noise, rather than the systematic image-text emotional conflicts characteristic of memes. How such adaptive fusion techniques can be leveraged to alleviate negative transfer remains underexplored.

2.4. Dataset examples and challenges

Before introducing the methodology, it is important to highlight the challenges of multimodal emotion recognition in social media datasets, particularly those containing memes. Memes often exhibit surface-level emotional cues (e.g., smiling faces) that contradict the underlying sentiment conveyed by the textual or shared cultural context. This discrepancy significantly increases the difficulty of emotion recognition compared with conventional multimodal posts.

As shown in Figure 1, both images depict smiling faces. In the first case, the image is a manipulated meme in which a laughing face is altered with an animal-like feature, conveying irony and ridicule, and therefore annotated as negative. In contrast, the second example is a conventional social media post showing people smiling in a genuine context, which is annotated as positive. Although both images share the same visual cue (a smile), their emotional implications are fundamentally different. This demonstrates why recognising emotions in meme-rich environments is particularly challenging for multimodal models.



(a) Trump meme with pig nose (ironic/negative).

(b) Conventional smiling photo (genuine/positive).

Figure 1. Challenge of multi-modal emotion recognition on meme-rich datasets

Despite similar visual cues (smiling faces), the underlying sentiments differ significantly.

This paper distinguishes itself from existing research by being the first to systematically analyse the performance degradation and negative transfer effect caused by memes in a multimodal emotion recognition framework. We not only quantify the actual impact of memes on emotion recognition, but also propose a dedicated framework that combines gated adaptive fusion with multi-task joint learning. By introducing meme detection as an auxiliary task and explicitly modelling image–text conflict through adaptive gating, our approach directly addresses the key gaps identified above.

3. Methodology

3.1. Data preprocessing

3.1.1. Data sources and composition

This study constructs a mixed dataset to simulate a real-world social media environment in which meme and non-meme content coexist. Data were sourced from two publicly available datasets:

- MVSA-Single: Contains 5,129 conventional social media image-text pairs annotated with three-class sentiment labels (positive, neutral, negative) [22].
- MET-Meme: A dedicated meme dataset comprising 3,994 English image-text pairs and labels, providing fine-grained emotion and metaphor annotations [23].

We randomly selected 4,000 image-text pairs with their labels from MVSA-Single and used all 3,994 image-text pairs with labels from MET-Meme to construct a balanced mixed dataset totalling 7,994 samples. The dataset was divided into training and test sets in an 8.5:1.5 ratio. After data cleaning (including low-quality text filtering, language screening, etc.), the final test set contained 887 valid samples, consisting of 401 meme posts and 486 regular image-text posts.

3.1.2. Emotion label normalisation

Due to differences in the annotation schemes of the two datasets, directly merging them would cause inconsistencies in model training and evaluation. MET-Meme employs a more fine-grained taxonomy (e.g., amusement, sorrow, anger, fear, with some samples also annotated for emotion intensity), whereas MVSA-Single uses a three-class scheme (positive, neutral, negative).

To standardise the labels and facilitate cross-dataset comparison, MET-Meme's categories were mapped into the same three sentiment classes as MVSA-Single:

- Positive: amusement, joy, love
- Neutral: original neutral labels and ambiguous or low-polarity categories
- Negative: anger, sorrow, disgust, fear

This mapping preserves emotional polarity while reducing fluctuations caused by differences in label distribution.

3.1.3. Low-quality post filtering

To improve the quality of text-image posts' features, a systematic low-quality post filtering process was applied:

- Texts shorter than six characters were considered invalid.
- Corrupted or damaged images/texts were treated as invalid.
- Texts with more than 70% non-alphanumeric characters were regarded as low-quality.
- Non-English texts (detected via langdetect) were replaced with empty strings to ensure compatibility with BERT-base-uncased.

All samples were retained; however, invalid text entries were replaced with empty strings. A new field, *has_comment*, was added (1 = valid text, 0 = empty).

3.2. Baseline model

The baseline model in this study follows an early fusion framework [24], one of the most widely adopted approaches in Multimodal Emotion Recognition (MER). The overall architecture is shown in Figure 2. The central idea is to extract features independently from the text and image modalities, concatenate them at the vector level, and then feed the combined representation into a single classifier for emotion prediction. Owing to its simplicity, low computational cost, and widespread use as a reference point, this framework remains a common baseline in multimodal research [25, 26].

In practice, the text modality is encoded using a pre-trained BERT model to produce a fixed-dimensional semantic vector. The image modality is processed with ResNet-18, pre-trained on ImageNet, to extract visual features, which are then projected into the same dimensional space as the text representation via a fully connected layer [26, 27]. The joint multimodal representation is obtained through feature concatenation, as illustrated in Equation (1):

$$h_{joint} = h_{text} \oplus h_{image} \tag{1}$$

where h_{text} and h_{image} denote the text and image feature vectors, respectively, and $\oplus$ represents concatenation. The joint representation h_{joint} is then passed through a fully connected layer and a softmax classifier to predict the emotion label (*positive, neutral*, or *negative*).

This baseline is consistent with widely used early fusion approaches in multimodal research, such as the Tensor Fusion Network and the official baseline in the Hateful Memes Challenge [13].

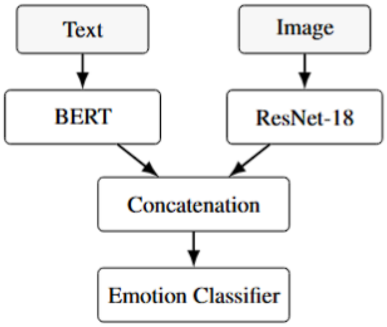


Figure 2. Baseline model architecture (early fusion by concatenation)

Using this model, therefore, provides a transparent and fair point of comparison for evaluating the proposed framework [24, 28].

3.3. Proposed method

This study proposes MAG-Net, a gated multi-task fusion framework. Built upon a unified multimodal encoding structure, the framework jointly optimises emotion classification and meme detection, and introduces an adaptive gated fusion mechanism together with a sample-weighting strategy based on meme probability, thereby significantly enhancing the model's robustness in meme-rich scenarios. The overall architecture is shown in Figure 3.

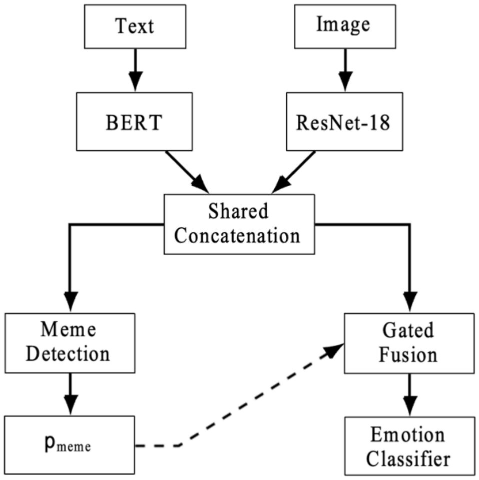


Figure 3. Proposed MAG-Net architecture (gated fusion with multi-task learning)

The shared concatenation layer feeds two branches: meme detection outputs pmeme, which serves as a gating signal to the fusion module.

3.3.1. Overall architecture

In the feature encoding stage, we adopt the same pre-trained feature extractors as the baseline model. Text inputs are processed using BERT-base-uncased, producing a 768-dimensional semantic feature vector h_{text}; while image inputs are encoded using ResNet-18 pre-trained on ImageNet, yielding a 512-dimensional visual feature vector h_{image}. Unlike the baseline model, which directly concatenates features for sentiment classification, the core innovation of MAG-Net lies in the introduction of a shared feature concatenation layer that serves as a multi-task information hub, together with lightweight task-specific heads that enable efficient multi-task collaboration.

As shown in Figure 3, the text and image features are first aligned to a unified dimension (d = 512) through independent linear projection layers, yielding $\widetilde{h}_{text}$ and $\widetilde{h}_{image}$. These aligned features are then concatenated to form a shared multimodal representation:

$$h_{concat} = [\widetilde{h}_{text} \oplus \widetilde{h}_{image}] \in R^{1024} \tag{2}$$

This concatenation layer (Shared Concatenation) serves as the central node of the entire architecture, providing feature input to two task branches simultaneously:

1. Left branch (meme detection)

A lightweight binary classification head is attached to h_{concat}. This head consists of only a single fully-connected layer, which directly outputs the meme probability $p_{meme} \in [0, 1]$ through a sigmoid activation function. This minimalistic design avoids over-parameterisation of the auxiliary task while ensuring effective gradient backpropagation to the shared encoding layers, promoting feature synergy between tasks.

2. Right branch (gated fusion and emotion classification)

h_{concat} together with p_{meme} generated from the left branch are fed into the gated fusion module. This module dynamically generates a weight vector g based on the meme probability, and performs adaptive weighted fusion of the aligned textual and visual features, outputting the final emotion classification feature h_{fused}. This feature is then passed to the emotion classification head for three-class prediction.

Throughout the multi-task framework, the BERT and ResNet-18 encoders as well as the feature concatenation layer are fully shared between the two tasks, while each task only performs output mapping through its own dedicated lightweight prediction head (a single fully-connected layer). This design ensures consistency in feature representation while minimising the parameter growth and computational overhead introduced by multi-task learning, achieving efficient and compact end-to-end joint learning. Through the combination of shared encoding layers and lightweight task heads, the model can simultaneously optimise emotion classification and meme detection during training, and dynamically mitigate cross-modal semantic conflicts caused by memes during inference via the gating mechanism.

3.3.2. Gated fusion module

Unlike the simple concatenation used in the baseline, the gated fusion module employs a learned weight vector $g \in [0, 1]^{512}$ to adaptively balance the contributions of text and image features. The generation of the gating weights depends not only on the concatenated multimodal feature h_{concat} but also incorporates the meme probability p_{meme} output by the auxiliary task:

$$g = \sigma W_g [h_{concat} \oplus p_{meme}] + b_g \tag{3}$$

where σ denotes the sigmoid function, $W_g \in R^{512\times1025}$ and $b_g \in R^{512}$ are the learnable parameters of the gating network, and $\oplus$ indicates vector concatenation.

The final fused representation is obtained by weighting the projected textual feature $\widetilde{h}_{text}$ and visual feature $\widetilde{h}_{image}$ with the gating weights:

$$h_{fused} = g \odot \widetilde{h}_{text} + (1 - g) \odot \widetilde{h}_{image} \tag{4}$$

where $\odot$ denotes element-wise multiplication.

This design enables the model to automatically reduce the influence of potentially misleading visual features when the meme probability p_{meme} is high, thereby effectively suppressing the cross-modal semantic conflicts common in memes and enhancing the model's robustness in mixed-content scenarios [29, 30].

3.3.3. Multi-task learning

The overall training objective of the model is defined as a weighted sum of the emotion classification loss and the meme detection loss, incorporating an adaptive weighting mechanism based on the meme probability:

$$L = L_{emotion} + \lambda \cdot p_{meme} \cdot L_{meme} \tag{5}$$

where $L_{emotion}$ is the cross-entropy loss for three-class (positive, neutral, negative) emotion recognition, L_{meme} is the binary cross-entropy loss for meme detection, and λ is a fixed hyper-parameter that controls the overall influence of the auxiliary task. The key innovation lies in introducing p_{meme} as a sample-wise adaptive weight, enabling the model to focus more on samples that are more likely to be memes during training, thereby enhancing its ability to perceive cross-modal semantic conflicts. Meme detection is formulated as an auxiliary binary classification task that shares the same BERT and ResNet-18 encoders as well as the feature concatenation layer with emotion classification. A lightweight classification head consisting of only a single fully-connected layer is applied to the concatenated features to predict the meme probability. During training, the meme detection head is optimised jointly with the emotion classifier, encouraging the shared encoding layers to learn feature representations useful for both tasks. At inference time, the predicted meme probability p_{meme} is fed as a gating signal into the fusion module to dynamically adjust the weighting of textual and visual features. In this way, meme detection not only enhances the model's generalisation as an auxiliary task but also directly participates in the decision-making process of cross-modal fusion, thereby significantly improving the overall robustness of the system in meme-rich scenarios [8, 31].

3.3.4. Implementation details

Gating Network: A single fully-connected layer takes the 1024-dimensional concatenated feature together with the 1-dimensional meme probability as input (total 1025 dimensions) and outputs a 512-dimensional gating weight vector g.

Meme detection head: A single fully-connected layer maps the 1024-dimensional concatenated feature to a 1-dimensional meme probability p_{meme}, followed by a sigmoid activation.

Training: The Adam optimiser is used with an initial learning rate of 1×10^{-4} and a batch size of 16. Early stopping is applied: training terminates if the validation loss does not improve for 5 consecutive epochs.

Hardware: Experiments are conducted on a single NVIDIA 4060 GPU.

4. Results

4.1. The failure and repair of multimodal fusion

We first compare the proposed gated multi-task framework against the standard early fusion baseline under different training and testing conditions. The key results are summarised in Table 1.

When trained exclusively on regular data, the early fusion model achieves a three-class emotion recognition accuracy of 0.6750 on the regular test set, which represents the performance upper bound for this task under ideal conditions. However, when trained on the mixed dataset (50% memes), the early fusion model's three-class emotion recognition accuracy drops to 0.4228. More critically, its performance on the regular subset plummets to 0.3704, demonstrating a severe negative transfer effect where meme samples degrade the recognition of regular content.

In contrast, our method effectively repairs this failure. Under the same mixed training, our gated multi-task model achieves 0.5693 overall emotion recognition accuracy, substantially outperforming the baseline. Most importantly, it fully restores and even surpasses performance on regular content, reaching 0.6893 accuracy on the regular subset. This confirms that our framework successfully isolates the disruptive influence of memes.

Table 1. Performance comparison: early fusion vs. our gated multi-task method

Training	Model	Test: Overall	Test: Meme Subset	Test: Normal Subset
Regular only	Early Fusion	-	-	0.6750
Mixed data	Early Fusion	0.4228	0.4863	0.3704
Mixed data	Proposed (Ours)	0.5693	0.4239	0.6893

4.2. Effectiveness of the proposed method

As shown in Table 2, the proposed model achieves an accuracy of 0.5693, representing a relative improvement of 34.6% over the early fusion baseline (0.4228). Consistent gains are observed across all macro-averaged metrics: precision improves from 0.4150 to 0.5720, recall from 0.4164 to 0.5680, and F1-score from 0.4070 to 0.5691.

Table 2. Overall performance on the mixed test set (887 samples)

Model	Acc	Macro-P	Macro-R	Macro-F1
Early Fusion	0.4228	0.4150	0.4164	0.4070
Proposed	0.5693	0.5720	0.5680	0.5691

4.3. In-depth analysis

4.3.1. Performance breakdown by content type

The performance breakdown in Table 3 highlights two key observations:

Restoration of Normal Content Performance: On the normal subset, our method achieves an accuracy of 0.6893 compared to the baseline's 0.3704. The neutral class shows strong improvement in recall (0.3200 -> 0.6600).

Stable Performance on Memes: On the meme subset, while the baseline shows slightly higher accuracy (0.4863 vs. 0.4239), both models achieve identical Macro-F1 scores (0.3851). The improved neutral recall on memes (0.1500 -> 0.2073) further indicates better handling of ironic expressions.

Table 3. Performance breakdown on meme and normal subsets

Subset	Model	Accuracy	Macro-F1	Neutral Recall
Meme Subset	Baseline	0.4863	0.3851	0.1500
(401 samples)	Proposed	0.4239	0.3851	0.2073
Normal Subset	Baseline	0.3704	0.3625	0.3200
(486 samples)	Proposed	0.6893	0.6865	0.6600

4.3.2. Auxiliary task performance

As shown in Table 4, the auxiliary meme detection task achieves high performance. This reliable detection provides the gating module with accurate input about content type.

Table 4. Performance of the auxiliary meme detection task

Class	Precision	Recall	F1	Samples
Normal	0.9391	0.9527	0.9459	486
Meme	0.9416	0.9252	0.9333	401
Macro Avg	0.9404	0.9389	0.9396	887

4.3.3. Error patterns

The confusion matrices show that the largest gains occur in the neutral and positive classes. For the baseline, 113 out of 274 neutral samples are misclassified as negative; our method reduces this to 81 misclassifications. The recognition of neutral memes remains difficult (recall: 0.2073).

4.3.4. Summary and discussion of results

Our gated multi-task framework not only achieves superior overall performance but also restores the model's ability to recognise normal social media content. Severe negative transfer occurs in traditional multimodal emotion recognition models in the presence of memes (normal content accuracy dropped from 0.6750 to 0.3704), and the proposed framework restores and increases normal content accuracy to 0.6893.

5. Discussion and conclusion

5.1. Summary of the study

This study is the first to systematically reveal and mitigate the negative transfer effect caused by memes in multimodal emotion recognition. By designing the MAG-Net, we achieve adaptive suppression of image–text semantic conflicts. While maintaining recognition performance on regular content, the framework significantly improves the overall robustness of the model in meme-rich scenarios. Experimental results demonstrate that the framework can effectively distinguish and handle conflict-prone samples, providing a new perspective for modelling multimodal conflicts.

5.2. Future directions

The current research is primarily based on a specific dataset and a three-class emotion taxonomy. Future work could be extended in the following directions:

1. Cross-cultural extension: Introducing more diverse meme-culture data to enhance the model's generalisation ability.
2. Fine-grained modelling: Extending to fine-grained emotion categories and incorporating explicit semantic understanding modules such as irony and metaphor detection.
3. Dynamic optimisation: Exploring adaptive loss weighting and more flexible multi-task balancing mechanisms.

This study offers a generalisable solution for conflict mitigation in multimodal learning and offers a technical reference for building emotion analysis systems in real-world social media scenarios.

References

[1] Handayani, F., Sari, S. D. S. R., & Respati, W. (2016). The use of meme as a representation of public opinion in social media. *Humaniora, 7*(3), 333–339.

[2] Bowo, F. A., Anisah, A., & Marthalia, L. (2024). Meme marketing: Generation Z consumer behavior on social media. *Jurnal Indonesia Sosial Sains, 5*(2), 188–201.

[3] Dupuis, M. J., & Williams, A. (2019). The spread of disinformation on the web: An examination of memes on social networking. In *2019 IEEE SmartWorld Conference* (pp. 1412–1418). IEEE.

[4] Sharma, S., Ramaneswaran, S., Akhtar, M. S., & Chakraborty, T. (2024). Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing, 15*(3), 1800–1811.

[5] Bejan, I. (2020). MemoSYS at SemEval-2020 task 8: Multimodal emotion analysis in memes. *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1172–1178). International Committee for Computational Linguistics.

[6] Routhu, R. K., & Baruah, U. (2025). Sentiment analysis on memes: A review. *Expert Systems, 42*(1), 1–21.

[7] Liang, B., Lou, C., Li, X., Yang, M., Gui, L., He, Y., Pei, W., & Xu, R. (2022). Multimodal sarcasm detection via cross-modal graph convolutional network. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 1767–1777). Association for Computational Linguistics.

[8] Wang, J., Yang, Y., Jiang, Y., Ma, M., Xie, Z., & Li, T. (2024). Cross-modal incongruity aligning and collaborating for multimodal sarcasm detection. *Information Fusion, 103*, 102132.

[9] Yao, X., She, D., Zhao, S., Liang, J., Lai, Y.-K., & Yang, J. (2019). Attention-aware polarity-sensitive embedding for affective image retrieval. *2019 IEEE/CVF International Conference on Computer Vision* (ICCV) (pp. 1140–1150). IEEE.

[10] Goel, P., & Vishwakarma, D. K. (2024). Impact of attention on visual sentiment analysis. *2024 14th International Conference on Cloud Computing, Data Science & Engineering* (Confluence) (pp. 703–707). IEEE.

[11] Li, H., Liu, H., Yu, P., Zhao, J., Wan, B., & Li, W. (2023). Deep learning-based approach for emotion recognition using image-text fusion. *2023 International Joint Conference on Information and Communication Engineering* (JCICE) (pp. 11–15). IEEE.

[12] Wan, X., Liang, J., & Zhang, H. (2025). *EmoHeal: An end-to-end system for personalized therapeutic music retrieval from fine-grained emotions*. arXiv. https: //arxiv.org/abs/2509.15986

[13] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2021). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems, 33*, 2611–2624. https: //proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html

[14] 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.

[15] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125.

[16] Jia, J., Zhang, H., & Liang, J. (2025). Bridging discrete and continuous: A multimodal strategy for complex emotion detection. *2025 IEEE International Workshop on Machine Learning for Signal Processing* (MLSP) (pp. 1–6). IEEE.

[17] Ramamoorthy, S., Gunti, N., Mishra, S., Suryavardan, S., Reganti, A. N., Patwa, P., Das, A., Chakraborty, T., Sheth, A. P., Ekbal, A., & Ahuja, C. (2022). Memotion 2: Dataset on sentiment and emotion analysis of memes. *Proceedings of the AAAI 2022 Workshop on De-Factify: Combating Online Multimodal Disinformation*.

[18] Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content. *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying* (pp.

32–41). European Language Resources Association.

[19] Liang, J., Liu, X., Wang, W., Plumbley, M. D., Phan, H., & Benetos, E. (2025). Acoustic prompt tuning: Empowering large language models with audition capabilities. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 33*, 949–961.

[20] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

[21] Rayatdoost, S., Rudrauf, D., & Soleymani, M. (2020). Multimodal gated information fusion for emotion recognition. *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 655–659). Association for Computing Machinery.

[22] Yu, J., & Jiang, J. (2019). Adapting BERT for target-oriented multimodal sentiment classification. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 5408–5414). International Joint Conferences on Artificial Intelligence Organization.

[23] Xu, B., Li, T., Zheng, J., Naseri-parsa, M., Zhao, Z., Lin, H., & Xia, F. (2022). MET-Meme: A multimodal meme dataset rich in metaphors. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2887–2899). Association for Computing Machinery.

[24] Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., & Aßenmacher, M. (2023). *Multimodal deep learning.* arXiv. https: //arxiv.org/abs/2301.04856

[25] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

[26] Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. *Proceedings of the 20th Chinese National Conference on Computational Linguistics* (pp. 1218–1227). Chinese Information Processing Society of China.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE.

[28] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1103–1114). Association for Computational Linguistics.

[29] Lee, K., Kim, J., Chong, S., & Shin, J. (2017). *Making stochastic neural networks from deterministic ones*. arXiv. https: //arxiv.org/abs/1704.03058

[30] Xue, Z., & Marculescu, R. (2023). *Dynamic multimodal fusion.* arXiv. https: //arxiv.org/abs/2204.00102

[31] Ruder, S. (2017). *An overview of multi-task learning in deep neural networks*. arXiv. https: //arxiv.org/abs/1706.05098