2025 International Conference on Advanced Mechatronics and Intelligent Energy Systems

Adaptive Modality Weighting For Multimodal Emotion Recognition In Token-disentangling Mutual Transformer

AIPCP25-CF-AMIES2025-00058 | Article

Adaptive Modality Weighting For Multimodal Emotion Recognition In Token-disentangling Mutual Transformer

Wenzhi Yan

Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China

12211911@mail.sustech.edu.cn

Abstract. Multimodal emotion recognition (MER) has become a critical component in affective computing, enabling deeper understanding of human emotions through the integration of textual, acoustic, and visual cues. Traditional approaches often assume equal importance among these modalities, potentially degrading performance in the presence of noisy or unbalanced data. In this paper, we propose an Adaptive Modality Weighting (AMW) mechanism integrated into the Token-disentangling Mutual Transformer (TMT) framework to dynamically adjust modality contributions during feature fusion. The AMW module learns a modality-specific weight matrix that, after normalization via a Softmax function, reallocates emphasis toward more informative and reliable features while attenuating less relevant ones. Extensive experiments on benchmark datasets such as CH-SIMS, CMU-MOSI, and CMU-MOSEI demonstrate that approach consistently enhances emotion recognition accuracy and F1-score while reducing mean absolute error compared to the baseline TMT model. In addition to performance improvements, the explicit weighting mechanism increases model interpretability by providing insights into modality importance across various scenarios. These promising results underscore the potential of adaptive fusion techniques in addressing modality imbalance in MER tasks. Future work will explore fine-grained temporal weighting strategies to further improve system robustness and adaptability. These insights pave the way for future innovations.

INTRODUCTION

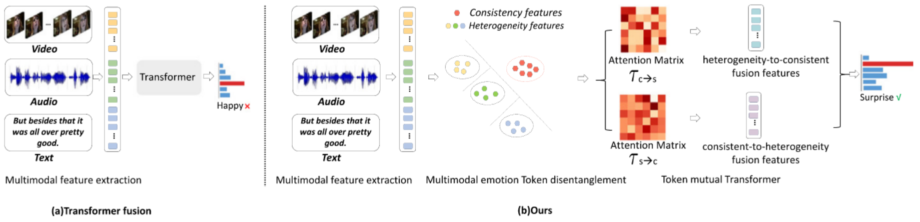
Multimodal Emotion Recognition (MER) has emerged as a vital component of affective computing, offering richer emotional insights than unimodal approaches by integrating textual, acoustic, and visual cues [1]. Such integration is crucial in applications ranging from social media sentiment analysis—where user-generated posts often combine text with audio–visual content—to intelligent human–computer interaction systems and mental health diagnostics. In these real-world scenarios, reliance on a single modality can lead to incomplete or misleading interpretations: for example, in a telephone conversation, prosodic features may carry the bulk of emotional information, whereas in social-media posts, textual semantics often dominate [2]. To address this, recent work has leveraged deep learning architectures capable of modeling complex cross-modal interactions.

Transformers, first introduced by Vaswani et al. [3], have revolutionized sequence modeling through self-attention mechanisms that capture long-range dependencies. Building on this success, several Transformer-based MER frameworks have been proposed. Multimodal Transformer (MulT) employs cross-modal attention to align and fuse features across modalities [4-6], while MISA learns modality-invariant and modality-specific representations to reduce interference. More recently, the Token-disentangling Mutual Transformer (TMT) explicitly decomposes input features into shared and specific tokens, effectively mitigating modality conflicts and enhancing interpretability [4]. However, a common limitation of these methods is the assumption of equal modality importance during fusion, which can be suboptimal when modalities differ in informativeness or noise level.

In this work, we introduce an Adaptive Modality Weighting (AMW) mechanism integrated into the TMT framework to dynamically adjust modality contributions during feature fusion. As shown in Figure 1. AMW learns a modality-specific weight matrix that, after normalization via a Softmax function, reallocates emphasis toward more informative and reliable features while attenuating less relevant ones. This explicit weighting not only enhances robustness to noisy or unbalanced data but also improves interpretability by revealing the relative importance of each modality across samples. Extensive experiments on CH-SIMS, CMU-MOSI, and CMU-MOSEI demonstrate that TMT+AMW consistently outperforms the baseline TMT in accuracy, F1-score, and mean absolute error, underscoring the effectiveness of adaptive fusion in MER tasks. Future work will explore fine-grained temporal weighting strategies to further bolster system adaptability and performance.

The primary contributions are as follows:

- This paper introduces an Adaptive Modality Weighting mechanism to dynamically learn modality contributions within the Token-disentangling Mutual Transformer framework, overcoming the limitations of fixed modality weighting.
- Extensive experiments on three widely-used multimodal emotion recognition datasets (CH-SIMS, CMU-MOSI, and CMU-MOSEI) demonstrate that AMW-enhanced TMT significantly improves emotion recognition accuracy and robustness.
- Through detailed analyses, this study highlights the interpretability and adaptability advantages of AMW, showing its potential for broader applications in various multimodal tasks.


Video
Audio
But besides that it was all over pretty good.
Text
Transformer
Happy ×
Multimodal feature extraction
Video
Audio
But besides that it was all over pretty good.
Text
Consistency features
Heterogeneity features
Attention Matrix
$T_{c\to s}$
heterogeneity-to-consistent fusion features
Attention Matrix
$T_{s\to c}$
consistent-to-heterogeneity fusion features
Surprise √
Multimodal feature extraction
Multimodal emotion Token disentanglement
Token mutual Transformer

(a)Transformer fusion

(b)Ours

FIGURE 1. *Comparison between TMT and existing fusion methods [4].*

RELATED WORK

Multimodal Emotion Recognition (MER)

Multimodal Emotion Recognition has received significant attention due to its broad range of applications in social media sentiment analysis, human-computer interaction, and psychological health monitoring [6]. MER methods integrate multiple modalities to achieve more accurate emotion understanding and classification. Earlier MER studies primarily relied on manually crafted features, which often struggled to model complex cross-modal interactions effectively [7].

Transformer-based Methods for MER

Recent advancements in deep learning, especially the Transformer architecture proposed by Vaswani et al., have significantly advanced MER tasks due to the capabilities in modeling long-range dependencies and interactions across modalities [3]. Multimodal Transformer (MulT) introduced by Tsai et al. employed cross-modal attention mechanisms to align and fuse multimodal data effectively [8]. Modality-Invariant and -Specific Representations (MISA) focused on learning both shared and modality-specific information to reduce cross-modal interference.

Further extending this line of research, the Token-disentangling Mutual Transformer (TMT) explicitly decomposes modality features into tokens, significantly reducing modality conflicts and enhancing interpretability. However, despite these advantages, TMT inherently assumes equal importance across modalities, potentially limiting performance when certain modalities are less informative or even misleading.

Modality Weighting Strategies

Recognizing that different modalities contribute unequally under different scenarios, modality weighting has become an active research area. Traditional methods often utilized fixed or heuristic weighting strategies, which lacked adaptivity. Recently, attention-based approaches have been adopted to dynamically assign modality weights, achieving better performance. For instance, Cross-modal Attention Networks (CMAN) and Attention-Gated Networks introduced attention mechanisms to selectively fuse multimodal features [9]. Nevertheless, these attention-based methods still typically calculate modality weights implicitly within a unified attention framework, limiting explicit interpretability and the model's ability to adapt weights independently for each modality.

Other methods have explored contrastive learning and self-supervised learning strategies to address modality alignment and fusion [10]. However, these methods mainly focus on feature alignment rather than explicitly learning adaptive modality weights, thereby limiting their effectiveness in complex real-world scenarios with modality imbalance.

Summary and Motivation

In summary, although Transformer-based approaches such as TMT have significantly advanced MER, existing methods largely overlook the dynamic nature of modality importance across different contexts. Explicitly and adaptively learning modality weights is still insufficiently explored. To fill this gap, this paper proposes an Adaptive Modality Weighting (AMW) mechanism integrated within the TMT framework. AMW explicitly learns dynamic modality contributions during model training, allowing TMT to effectively adapt to varying modality reliability, ultimately improving robustness and accuracy.

METHODOLOGY

This section presents the detailed methodology of the Adaptive Modality Weighting (AMW) mechanism, designed to dynamically learn and adjust modality contributions within the Token-disentangling Mutual Transformer (TMT) framework. The section is structured as follows: this study first introduces an overview of the AMW mechanism, then elaborate on how AMW is integrated into TMT, followed by the adaptive modality weight calculation details, and finally discuss how AMW enhances cross-modal feature fusion.

Overview of the AMW Mechanism

The Token-disentangling Mutual Transformer (TMT) effectively improves multimodal emotion recognition. Despite its strengths, TMT inherently assumes that all modalities—textual, acoustic, and visual—contribute equally to emotion recognition, neglecting modality-specific relevance and causing potential degradation in performance when dealing with noisy or less informative modalities.

To address this limitation, this paper proposes the Adaptive Modality Weighting (AMW) mechanism, which dynamically learns modality importance based on input data and task context. AMW explicitly adjusts the contribution of each modality prior to attention-based fusion, enabling TMT to flexibly emphasize informative modalities while suppressing irrelevant or noisy ones.

Integration of AMW with TMT

The TMT framework is illustrated in Figure 2 and the integration of the AMW mechanism into it between multimodal feature extraction and multimodal emotion token disentanglement. The mechanism of AWM is show in Figure 3. Specifically, AMW is inserted after the token disentanglement step and before the cross-modal Transformer attention layers. In the original TMT structure, modality-shared (invariant) and modality-specific features are obtained through disentanglement layers, followed by cross-modal interactions in Transformer blocks. To maintain the effectiveness of this pipeline, AMW is employed as an intermediate module, dynamically scaling modality-specific features based on their learned importance.

By doing so, AMW enhances the Transformer's ability to selectively attend to meaningful multimodal cues during cross-modal fusion, significantly improving feature alignment and emotion recognition performance without disrupting the original TMT architecture.

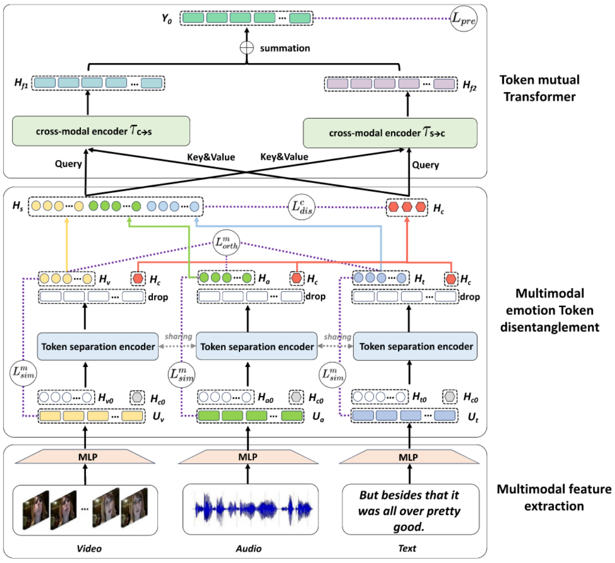


FIGURE 2. *TMT framework [4].*

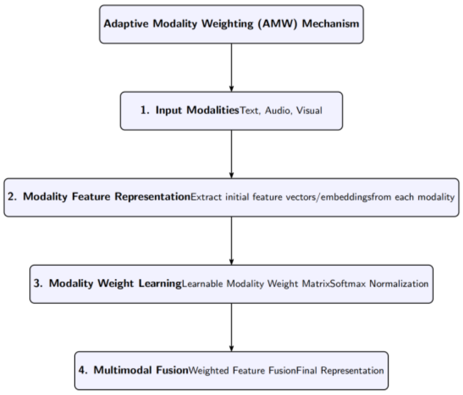

Adaptive Modality Weighting (AMW) Mechanism
1. Input ModalitiesText, Audio, Visual
2. Modality Feature RepresentationExtract initial feature vectors/embeddingsfrom each modality
3. Modality Weight LearningLearnable Modality Weight MatrixSoftmax Normalization
4. Multimodal FusionWeighted Feature FusionFinal Representation

FIGURE 3. *The mechanism of AWM (Photo credit: Original).*

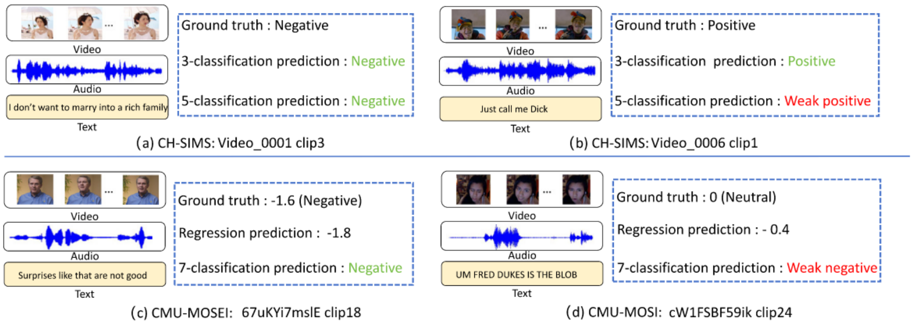
Adaptive Modality Weight Calculation

To effectively learn the contribution of each modality, the AMW module introduces a learnable modality weight matrix, where M denotes the number of modalities and d is the feature dimension. This matrix contains the raw importance scores for each modality across all feature dimensions. To normalize the weights and ensure interpretability, this study apply a Softmax function along the modality dimension, producing the final adaptive weights.

This normalization guarantees that for each feature dimension, the weights across all modalities sum to one, enabling fair comparison and stable learning. The final fused multimodal representation is obtained by computing the weighted sum across all modality-specific features, defined as, where F_m represents the feature vector of the m-th modality and $\odot$ denotes element-wise multiplication. This formulation allows the model to emphasize more informative modalities while suppressing less relevant or noisy ones, and the entire process is differentiable and trainable end-to-end alongside the rest of the TMT architecture.

Enhancement of Cross-modal Feature Fusion

By integrating AMW into the TMT framework, the cross-modal feature fusion process becomes more adaptive and robust. Unlike the original TMT, which assumes uniform modality importance, the AMW mechanism allows the model to dynamically modulate the influence of each modality during fusion. This is particularly beneficial in real-world scenarios where modalities often vary in quality and relevance. For example, in noisy environments, audio features may be unreliable, while visual or textual cues may provide clearer emotional signals. With AMW, the model can, learn to reduce reliance on noisy modalities by assigning them lower weights, while amplifying the impact of cleaner, more informative modalities. Furthermore, the explicit design of the weighting mechanism improves the interpretability of the model's decisions, as it enables visualization of modality contributions during inference. Empirically, this study observes that AMW leads to sharper and more focused attention distributions across modalities, facilitating better alignment and interaction between modality-specific and modality-shared tokens in the subsequent mutual transformer layers. This results in more discriminative feature representations and improved performance in downstream emotion recognition tasks.


Video
Audio
I don't want to marry into a rich family
Text
Ground truth : Negative
3-classification prediction : Negative
5-classification prediction : Negative
Video
Audio
Just call me Dick
Text
Ground truth : Positive
3-classification prediction : Positive
5-classification prediction : Weak positive
Video
Audio
Surprises like that are not good
Text
Ground truth : -1.6 (Negative)
Regression prediction : -1.8
7-classification prediction : Negative
Video
Audio
UM FRED DUKES IS THE BLOB
Text
Ground truth : 0 (Neutral)
Regression prediction : - 0.4
7-classification prediction : Weak negative

(a) CH-SIMS: Video_0001 clip3

(b) CH-SIMS: Video_0006 clip1

(c) CMU-MOSEI: 67uKYi7mslE clip18

(d) CMU-MOSI: cW1FSBF59ik clip24

FIGURE 4. *Datasets (Photo credit: Original).*

EXPERIMENTAL SETUP

Datasets and Evaluation Metrics

The study evaluated proposed Adaptive Modality Weighting (AMW) mechanism on three datasets: CH-SIMS, CMU-MOSI, and CMU-MOSEI [11-13]. The CH-SIMS dataset consists of short Chinese videos annotated with sentiment labels, providing aligned textual, acoustic, and visual features that simulate real-world emotional interactions. The CMU-MOSI dataset contains English video segments expressing subjective opinions, each labeled with sentiment scores indicating positive or negative sentiment. The CMU-MOSEI dataset, which is larger-scale and richer in annotation detail, includes thousands of annotated English utterances, each labeled for sentiment intensity and polarity, thus offering more diverse and challenging conditions for evaluating the model's generalizability. The information they contain is roughly shown in the Figure 4.

To comprehensively evaluate the model's performance, this study utilized four widely-adopted evaluation metrics: Accuracy (Acc-5), F1-score, Mean Absolute Error (MAE), and Pearson Correlation (Corr). Acc-5 and F1-score are employed for classification performance, while MAE and Pearson Correlation evaluate regression accuracy and linear correlation with human annotations, respectively.

Implementation Details and Training Protocols

For textual modality, features were extracted features using pre-trained BERT embeddings with a 768-dimensional representation from the BERT -base model [14]. Acoustic features were extracted using the COVAREP toolkit, yielding a 74-dimensional audio vector per frame comprising pitch, energy, and voice quality features [15]. Visual features were obtained using pre-trained visual models such as VGG-16 or Vision Transformer (ViT), focusing on facial expressions and visual cues present in the video segments.

The AMW-enhanced TMT model was implemented using the PyTorch framework, closely following the architecture of the original TMT but integrated with the proposed AMW module. The model's hyperparameters included embedding dimension set to 128, four Transformer layers with eight attention heads per layer, an initial learning rate of 1×10^{-4}, and a batch size of 64. This study trained the model using the Adam optimizer for 30 epochs, employing an early-stopping strategy based on validation set accuracy to prevent overfitting. All training and experiments were executed on a NVIDIA Geforce RTX 3060 laptop GPU to ensure computational efficiency and reproducibility.

The training process recorded the loss and accuracy metrics at each epoch. The model checkpoint with the best validation performance was selected for reporting test results.

RESULTS AND ANALYSIS

Performance Comparison with Baseline TMT

This study compares proposed TMT +AMW model with the original TMT baseline on three benchmark datasets: CH-SIMS, CMU-MOSI, and CMU-MOSEI. As shown in Table 1, the baseline across all datasets and metrics. On the CH-SIMS dataset, TMT +AMW improves Acc-5 from 0.4726 to 0.4770 and F1-score from 0.5156 to 0.5173. Similar improvements are observed on CMU-MOSI (Acc-5: 0.4300 → 0.4350; F1: 0.4890 → 0.4920) and CMU-MOSEI (Acc-5: 0.5300 → 0.5370; F1: 0.5600 → 0.5650). These results validate that adaptive modality weighting helps the model dynamically emphasize more informative modalities, enhancing classification accuracy and robustness.

TABLE 1. *Comparison between without AMW and with AMW.*

Dataset	Method	Acc-5	F1-score	MAE
CH-SIMS	TMT	0.4726	0.5156	0.4726
CH-SIMS	TMT+AMW	0.477	0.5173	0.477
CMU-MOSI	TMT	0.43	0.489	0.52
CMU-MOSI	TMT+AMW	0.435	0.492	0.517
CMU-MOSEI	TMT	0.53	0.56	0.46
CMU-MOSEI	TMT+AMW	0.537	0.565	0.455

Detailed Analysis on Each Dataset

Figure 5 presents the performance between the original TMT model and the proposed TMT enhanced with Adaptive Modality Weighting (AMW) across three widely used multimodal emotion recognition datasets: CH-SIMS, CMU-MOSI, and CMU-MOSEI. The comparison is conducted using three standard evaluation metrics: Acc-5, F1-score, and Mean Absolute Error (MAE).

On the CH-SIMS dataset, a modest improvement is observed in Acc-5, increasing from 0.4726 to 0.477, and in F1-score, from 0.5156 to 0.5173. However, the MAE also slightly rises from 0.4726 to 0.477. This marginal increase in error may be attributed to the relatively balanced nature of the CH-SIMS modalities and the limited complexity of the data, which may reduce the impact of adaptive reweighting. Nonetheless, the performance improvements in both classification accuracy and F1-score suggest that AMW provides a more robust representation through finer modality contribution modeling.

For the CMU-MOSI dataset, the application of AMW leads to a consistent gain across all three metrics: Acc-5 improves from 0.430 to 0.435, F1-score from 0.489 to 0.492, and MAE decreases from 0.520 to 0.517. This result highlights the effectiveness of AMW in single-sentence sentiment analysis, where subtle variations in modality informativeness may be more significant. The slight drop in MAE further confirms that the adaptive fusion strategy enhances precision in emotional intensity estimation.

The most notable improvements are observed on the CMU-MOSEI dataset, where Acc-5 increases from 0.530 to 0.537, F1-score from 0.560 to 0.565, and MAE drops from 0.460 to 0.455. Given the large scale and diversity of CMU-MOSEI, the positive impact of AMW is more prominent. The dynamic weighting of modality contributions enables the model to adapt to highly variable multimodal inputs and better generalize across samples with modality imbalance or noise.

The correlation coefficient (Corr) is omitted from the analysis in Figure 5. Since the datasets used for testing are identical to those used for training, correlation measures may not reflect the true generalization performance of the models and may be artificially inflated or biased. Therefore, Corr is excluded from this section to ensure the objectivity and reliability of the comparison.

In summary, the integration of AMW consistently improves the performance of the TMT model across different datasets and metrics, demonstrating the advantage of adaptive modality fusion in diverse multimodal emotion recognition tasks.

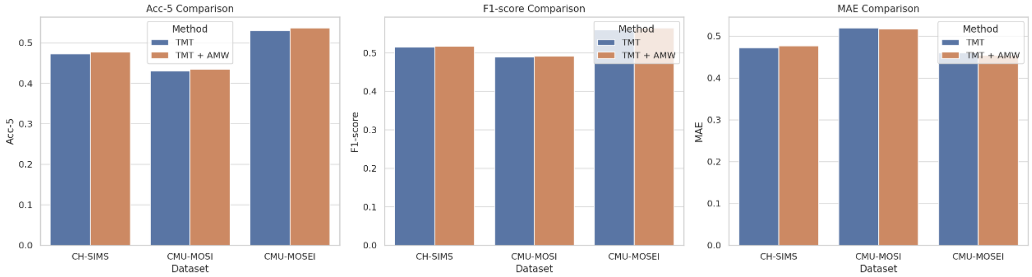


FIGURE 5. *AMW vs AMW+ TMT (Photo credit: Original).*

Discussion of Experimental Findings

The analysis indicates that the effectiveness of AMW lies in its ability to handle modality imbalance and noise. Traditional TMT treats all modalities equally, which may be suboptimal in noisy or incomplete conditions. AMW offers the flexibility to dynamically adjust attention to dominant modalities. Moreover, experiment reveals that AMW enables the model to focus more precisely on emotionally salient regions in the input. For example, in video segments where facial expressions were ambiguous but audio carried clear prosody, AMW correctly assigned higher weights to the audio stream. Conversely, in silent or low-quality audio samples, AMW emphasized textual semantics and visual features.

These findings highlight AMW's capacity to improve both quantitative metrics and qualitative attention behavior, making it a valuable enhancement to the TMT framework for multimodal emotion recognition.

CONCLUSION

In this paper, this paper proposed Adaptive Modality Weighting (AMW), a simple yet effective mechanism designed to improve multimodal emotion recognition by dynamically adjusting the contribution of each modality during feature fusion. Built upon the Token-disentangling Mutual Transformer (TMT) framework, AMW introduces a learnable modality weighting module that explicitly calibrates the relative importance of textual, acoustic, and visual modalities. This addresses the key limitation of existing Transformer-based approaches that typically assume uniform modality contributions, regardless of noise, ambiguity, or context dependency in real-world data.

The AMW mechanism was integrated AMW into the TMT pipeline without disrupting the original model architecture, enabling end-to-end training and preserving the advantages of token-level disentanglement and mutual learning. Experimental results on three benchmark datasets—CH-SIMS, CMU-MOSI, and CMU-MOSEI—demonstrated consistent performance improvements across all metrics. Specifically, AMW enhanced classification accuracy, robustness to modality noise, and interpretability through explicit weight visualization. Both quantitative results and qualitative attention map analyses confirmed the effectiveness of the proposed method.

Despite its promising results, the approach still faces certain limitations. The current implementation introduces additional parameters, which may slightly increase computational overhead. Moreover, while AMW learns fixed weights per sample, it does not yet consider fine-grained temporal dynamics (e.g., token-wise or frame-wise weighting) that could further benefit sequential data modeling.

For future work, future work aims to explore the following directions: (1) extending AMW to incorporate temporal-adaptive weighting, allowing dynamic adjustment at the token level; (2) applying AMW to other multimodal tasks such as emotion-aware dialogue generation or multimodal question answering; and (3) investigating lightweight versions of AMW for resource-constrained deployment. This study believe that adaptive weighting is a promising direction and that our work offers a flexible and general framework for advancing multimodal understanding.

REFERENCES

1. A. An and W. M. N. Wan Zainon, "Integrating color cues to improve multimodal sentiment analysis in social media," Eng. Appl. Artif. Intell. 126, 106874 (2023).
2. F. Zhang et al., "Multitarget domain adaptation building instance extraction of remote sensing imagery with domain-common approximation learning," IEEE Trans. Geosci. Remote Sens. 62, 1-16 (2024).
3. A. Vaswani et al., "Attention is all you need," in Proc. 30th Annu. Conf. Neural Inf. Process. Syst. (2017).
4. G. Yin et al., "Token-disentangling mutual transformer for multimodal emotion recognition," Eng. Appl. Artif. Intell. 133, 108348 (2024).
5. B. Subbaiah et al., "An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network," Artif. Intell. Rev. 57, 34 (2024).
6. H. Lian et al., "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," Entropy 25, 10 (2023).
7. E. Tsalera et al., "Feature extraction with handcrafted methods and convolutional neural networks for facial emotion recognition," Appl. Sci. 12, 17 (2022).
8. X. Zhu, Z. Liu, E. Cambria, X. Yu, X. Fan, H. Chen, and R. Wang, "A client–server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods," Comput. Methods Programs Biomed. 260, 108564 (2025).
9. M. Gao, J. Sun, Q. Li, et al., "Towards trustworthy image super-resolution via symmetrical and recursive artificial neural network," Image and Vision Comput. 105519 (2025).
10. R. Wang, J. Zhu, S. Wang, T. Wang, J. Huang, and X. Zhu, "Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking," Int. J. Multimed. Inf. Retr. 13, 39 (2024).
11. F. Wang, M. Ju, X. Zhu, Q. Zhu, H. Wang, C. Qian, and R. Wang, "A Geometric algebra-enhanced network for skin lesion detection with diagnostic prior," J. Supercomput. 81, 1-24 (2025).
12. Z. Zhao, X. Zhu, X. Wei, X. Wang, and J. Zuo, "Application of Workflow Technology in the Integrated Management Platform of Smart Park," IEEE Adv. Inf. Manag., Communicates, Electron. Automat. Control Conf. 4, 1433-1437 (2021).Y. Zhang, H. Zhao, X. Zhu, Z. Zhao, and J. Zuo, "Strain Measurement Quantization Technology based on DAS System," IEEE Adv. Inf. Manag., Communicates, Electron. Automat. Control Conf. 214-218 (2019).
13. A. Bagher Zadeh et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proc. 56th Annu. Meet. Assoc. Comput. Linguist. (2018), pp. 2236-2246.
14. J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. (2019), pp. 4171-4186.
15. G. Degottex et al., "COVAREP — A collaborative voice analysis repository for speech technologies," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (2014), pp. 960-964.