Attention Mechanism in Multimodal Emotion Recognition From 2020 to 2025: Technological Evolution, Challenges, and Future Prospects

Libin Han [1], Yu Huang [2], Chaoqian Liang [3, *] and Chao Sheng [4]

[1] Engineering Computer Science and Technology (Language Intelligence and Technology), Beijing Language and Culture University, Beijing, China

[2] English and Applied Linguistics, University of Nottingham China campus, Ningbo, Zhejiang, China

[3] Computer science and Technology, Jinan University, GuangZhou, GuangDong, China

[4] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China

* Corresponding Author Email: mrfurious666@stu2023.jnu.edu.cn

Abstract. Multimodal emotion recognition, as the core technology of human-computer interaction, has made breakthrough progress in the fusion of attention mechanisms in recent years. This article systematically reviews the research context in this field from 2020 to 2025, and for the first time establishes a classification framework from the dual dimensions of "attention refinement" and "modality interaction deepening". This paper will propose a four stage evolutionary model, including the process from foundation, to optimization, to multi-scale, and finally to cross fusion. It will also present a technical matrix for innovative induction of cross modal, recursive, multi-scale, and dynamic attention mechanisms. Furthermore, it will reveal the bottleneck of feature alignment caused by modal heterogeneity. Finally, a development path will be proposed from the directions of meta learning regulation, interpretability enhancement, etc., providing theoretical reference for constructing robust multimodal systems. Overall, this review not only summarizes the current progress of multimodal attention research but also highlights future opportunities for advancing human-computer interaction, encouraging continued exploration of more intelligent, inclusive, and adaptive multimodal systems.

Keywords: Multimodality, Emotion Recognition, Attention Mechanisms, Natural Language Processing, Large-scale Models.

1. Introduction

Emotion recognition aims to identify and interpret human emotions from diverse inputs, which is an increasingly important research area in artificial intelligence (AI) and human–computer interaction (HCI). Accurate emotion recognition is essential in social interactions, since human emotions are rich and multilayered, playing a critical role in shaping human behaviour [1,2]. Consequently, no single modality is sufficient to capture the full spectrum of emotional expression.

Traditional approaches to emotion recognition initially relied on unimodal inputs, such as analysing only text or speech. However, it is not possible to determine a person's emotional state solely from observing a specific entity or event [3,4]. This is one of the reasons why these early recognition systems frequently struggled in complex real-world environments.

This limitation gave rise to multimodal emotion recognition (MER), which combines multiple sources of information such as text, speech, and visual cues [5]. For example, in a conversation, an individual might say "I'm fine" with a trembling voice and tearful eyes. The words themselves suggest positivity, but the vocal and visual signals reveal distress. Multimodal systems aim to resolve such contradictions by integrating complementary evidence across modalities.

With recent advancements in architecture, deep learning has been extensively applied to MER [6]. A key technique for this task is the attention mechanism, first popularised in natural language processing. Attention enables models to dynamically assign weights to intra- or inter-modal features most relevant for emotion recognition—whether specific words, intonation, or fleeting facial

expressions [7]. For instance, in analysing text alongside video, an attention-based system can simultaneously detect frowns while highlighting negative vocabulary, giving these signals greater importance and suppressing irrelevant background noise. This capacity to prioritise context-specific features aids in overcoming noise, redundancy, and information imbalance inherent in multimodal data. In recent years, refinements in attention and the deepening of cross-modal interactions, alongside the introduction of large language models and reinforcement learning, have driven new progress in the interpretability and adaptability of MER.

This review will systematically examine developments in multimodal emotion recognition between 2020 and 2025, with particular emphasis on how attention mechanisms and fusion strategies have reshaped the field. By tracing its historical evolution, analysing the role of attention in recognition and fusion, and comparing leading models, this paper aims not only to highlight achievements but also to identify remaining challenges. Ultimately, it seeks to provide researchers, practitioners, and interdisciplinary collaborators with a clear perspective on progress in this field and potential directions for future innovation.

2. Development History of Multimodal Emotion Recognition

Over the past decade, multimodal emotion recognition has undergone significant transformation. Its development has consistently revolved around two central directions: the refinement of attention mechanisms and the deepening of modality interaction. Based on technological characteristics and chronological progression, four major stages can be distinguished: early foundational models, intermediate optimised models, multi-scale models, and cutting-edge cross-fusion models.

2.1. Early Foundational Models (2017–2019)

The earliest multimodal systems primarily relied on simple concatenation of modal features, such as combining speech and text embeddings. While straightforward, these approaches struggled to integrate heterogeneous features at varying temporal and semantic granularities. Representative work from this period, such as the TEMMA model (MuSe2021), employed multimodal multi-head attention (MMA) and temporal multi-head attention (TMA) to separately handle cross-modal interactions and temporal dependencies. However, its fixed-granularity utterance-level fusion was limited in performance, particularly in long-video contexts.

Another notable example is transformer-based models such as MulT, which introduced cross-attention to align text–audio–visual modalities. These models laid essential groundwork but only supported token-level, single-scale processing, meaning they could not effectively bridge fine-grained word-level features with coarser video-frame features.

2.2. Intermediate Optimised Models (2020–2022)

The next stage saw the introduction of recursive and dynamic weighting mechanisms, which allowed richer interactions across modalities. The Recursive Joint Cross-Modal Attention (RJCMA) framework exemplified this phase, iteratively refining associations between audio, text, and video features. Instead of processing input once, it revisited features across multiple rounds of attention, progressively strengthening emotional cues. On the Affwild2 dataset, RJCMA demonstrated significant improvements in arousal prediction [8].

Dynamic modality weighting also emerged as a key innovation. Frameworks such as Knowledge-guided Dynamic Attention (KuDA) adjusted each modality's contribution according to context. For example, when facial cues dominated, visual inputs were amplified. Such adaptive strategies helped to overcome the text-centred biases that constrained earlier fusion techniques.

Other RNN-based hybrid attention models, such as AM-RNN (2021), showed promise in dialogue analysis. Here, speaker states and contextual relations were modelled through bidirectional GRUs combined with attention [9]. This brought MER systems closer to real-time applicability in conversational systems. Paired with pairwise attention mechanisms, AM-RNN achieved 81.29% accuracy on the IEMOCAP dataset—a 2% improvement over conventional RNN approaches.

2.3. Multi-Scale Models (2023–2024)

As researchers recognised the limitations of fixed-granularity representations, focus shifted to multi-scale feature alignment. For instance, ScaleVLAD introduced a module capable of synchronously aligning text at word- and sentence-levels with video at frame- and segment-levels [10]. By incorporating a self-supervised clustering loss (S3C loss), the model enhanced feature discrimination and achieved state-of-the-art results on benchmark datasets. This allowed MER systems to capture both subtle local variations (e.g., sudden changes in tone) and broader emotional contexts (e.g., overall mood in a conversation).

Recursive multi-scale designs demonstrated that robust multimodal learning requires both fine-grained detail and holistic context. These approaches also underscored the importance of feature diversity in overcoming noise and redundancy in multimodal data.

2.4. Cutting-Edge Cross-Fusion Models (2024–2025)

The most recent stage has been characterised by cross-fusion architectures, enabling bidirectional information flow between modalities and integrating reinforcement learning with adaptive focus. A key example is the MemoCMT model, which employed cross-modal Transformers to establish bidirectional attention between text and audio. MemoCMT introduced a novel minimal aggregation strategy to highlight critical emotional cues while suppressing noise, achieving over 91% accuracy on the ESD dataset [11]. Similarly, CENet (2025) reduced inter-modal distributional differences through feature transformation strategies, outperforming state-of-the-art models across CMU-MOSI/MOSEI datasets. Its core relied on implicit dynamic adjustment mechanisms enabled by non-text vector indexing.

Likewise, R1-Omni integrated reinforcement learning into MER by using reward signals to guide attention allocation. This approach explicitly linked predictions with modal contributions (e.g., emphasising rapid tonal shifts combined with negative words as evidence of frustration), thereby enhancing model interpretability [12]. Such interpretability is particularly vital for applications in healthcare and education, where transparency in AI decision-making is indispensable.

3. Technical Analysis

3.1. Attention Mechanism in Multimodal Emotion Recognition

In Multimodal Emotion Recognition, the attention mechanism can achieve selective fusion of modal information. The attention mechanism can be classified into the following four typical types according to their structure and functional objectives.

Cross-modal Attention is a mechanism that enables different modalities of information to "communicate". It enables one modality to actively find the most relevant information in another modality, achieving deep integration and collaborative understanding eventually. The study shows that the MMA module of TEMMA achieves a CCC value of 0.5781 by calculating the audio-visual cross-weight [7]. Also, Bidirectional Cross-Attention in the MemoCMT allows audio to serve as the Query for retrieving text Keys, while text can also query audio features in the reverse direction. That improves the recognition rate by 3.2% to 5.7% compared with the unidirectional architecture [11]. And the pair-wise attention of AM-RNN explicitly constructs a text-audio-visual bimodal interaction matrix, improving the F1 value by 1-3% on CMU-MOSEI [9].

Recursive Cross-modal Attention solves the problem of insufficient features caused by single attention by optimizing the attention weights through multiple rounds of iteration. In Multimodal Emotion Recognition, especially in a video, the first round of attention will focus on the text to make an initial judgment of the emotion. The second and the third round will focus on voice and vision, paying attention to the tone, eye contact and micro-expressions.

After several rounds of iteration, the model is closer to the real emotional state. The study shows that RJCMA applies with a three-level recursive architecture. The first round focuses on the audio-visual temporal correlation, and subsequent rounds gradually incorporate textual semantics. The arousal CCC is improved by 9.2% compared to single attention on the Affwild2 dataset eventually [8].

Multi-Scale Attention's core breakthrough lies in its ability to adapt to heterogeneous granular features. In the Multimodal Emotion Recognition, that enables the model to adaptively select information across different granularities and modalities, instead of being "fixed granularity", therefore enhancing the robustness. The innovation of the VLAD module in the SclaeVLAD lies in processing word-level, sentence-level(text), frame-level and segment-level (video) features simultaneously. It aligns text phrases (such as "really good") as a whole with video segment-level features, achieving binary accuracy rates of 87.2/89.3 on the MOSI dataset [10].

The Dynamic Adaptive Attention Mechanism can automatically adjust the attention weights based on the characteristics of the input content, the intensity of emotions, or the context of the task. It is essential because the importance of features from different modalities varies in different multimodal sentiment analysis tasks. Some people have expressive faces, and in such cases, the attention weight of the image will be relatively high. The study shows that KuDA dynamically selects the dominant modality. In the visually dominant samples of the CH-SIMSv2 dataset, the visual weight is increased from the baseline 30% to 55%, improving Acc-5 by 8.3% [13]. Also, CENet's feature transformation strategy achieves implicit dynamic adjustment through the indexing of non-textual vectors, significantly optimizing the distribution of text representations in pre-trained language models [14].

3.2. Attention Mechanisms in Modality Fusion Strategies

The essence of modality fusion strategies lies in balancing complementarity and redundancy through attention mechanisms. Existing technical approaches can be categorized into four typical paradigms.

Early Fusion Strategies perform attention alignment at the feature level: The TEMMA model employs feature concatenation combined with a Temporal Multi-head Attention (TMA) module to process synchronized modality data. However, high-dimensional features are prone to noise interference. This model validated the effectiveness of early fusion in capturing temporal modality correlations on MuSe-Stress and MuSe-Physio tasks [7]. The improvement of ScaleVLAD lies in its multi-scale early fusion—aligning textual word-level and video segment-level features via shared semantic vectors. Its multi-scale VLAD module concurrently processes heterogeneous features of varying granularity, effectively mitigating the dimensionality curse in traditional early fusion. Experimental results on datasets like IEMOCAP and CMU-MOSI demonstrate the superiority of this strategy in adapting to modality heterogeneity [10].

Model-Level Fusion relies on attention iteration to optimize feature representation: The RJCMA model uses a recursive module to progressively refine weight allocation through three rounds of attention iteration. The first round focuses on audio-visual temporal correlations, while subsequent iterations gradually integrate textual semantics. This progressive optimization significantly improves arousal recognition performance in long videos from the Affwild2 dataset, achieving a CCC value of 0.619 on its test set [8]. ScaleVLAD adopts a multi-scale iteration mechanism to synchronously optimize local and global feature weights during VLAD clustering. Combined with a self-supervised shift-clustering loss (S3C Loss), it enhances the discriminability of fused features, achieving a 2.1% F1-score improvement over single-scale models on the CMU-MOSEI dataset [7]. Such methods excel

in scenarios like dialogues and long videos due to their recursive mechanisms for progressive feature refinement.

Cross-Attention Fusion establishes bidirectional semantic associations via Query-Key-Value structures: The Cross-Modal Transformer (CMT) module in MemoCMT constructs bidirectional interaction pathways for audio to text and text to audio. Experiments comparing MIN, MEAN, and CLS aggregation strategies reveal that MIN aggregation (highlighting key emotional cues) exhibits the strongest robustness against noisy modalities , achieving 91.93% UW-Acc on the ESD dataset—a 7.7% improvement over CLS aggregation [11]. Additionally, bidirectional attention weight calculation effectively captures implicit inter-modal semantic relationships, avoiding information loss caused by unidirectional attention.

Reinforced Fusion innovatively introduces task objectives to guide attention: The R1-Omni model employs a Reinforcement Learning with Verifiable Reward (RLVR) mechanism, using emotion recognition accuracy as a reward function to drive the attention module in dynamically filtering task-relevant cues. For instance, in user satisfaction prediction scenarios, the model automatically strengthens the association weights between rapid tone and negative vocabulary. This task-driven attention adjustment enables remarkable generalization when transferring from movie clip datasets to real-world dialogue datasets [12].

4. Multimodal Emotion Recognition: Performance Analysis of Existing Models and Future Trends

4.1. Comparison of Experimental Results and Performance of Existing Models

Table 1. Performance comparison of multimodal sentiment analysis models

Model	Dataset	Core Metric	Performance	References
AM-RNN	CMU-MOSEID	Accuracy	81.29%	[9]
TEMMA	MuSe-Stress	Valence CCC	0.6648 (Test Set)	[7]
ScaleVLAD	IEMOCAP	Average Acc	82.6%	[10]
RJCMA	Affwild2	Arousal CCC	0.619 (Test Set)	[8]
MemoCMT	ESD	UW-Acc	91.93%	[11]
CENet	CMU-MOSI	F1-score	88.63	[14]

The research evolution in multimodal emotion recognition is clearly reflected in the core improvements of various models, as summarized in Table 1. AM-RNN [9] pioneered the field by integrating speaker state GRU and a pair-wise attention mechanism to capture conversational context and participant interaction. TEMMA [7] then focused on early fusion and temporal attention, emphasizing the importance of capturing the dynamic changes in emotion from the beginning of data processing. In the same year, ScaleVLAD [10] demonstrated the value of extracting features from different time scales through multi-scale fusion and an optimized loss function.

Moving into 2024, RJCMA [8] deepened cross-modal interaction with a recursive cross-modal attention approach. The most recent models, MemoCMT [11] and CENet [14], represent current research trends. The former utilizes a bidirectional cross-attention to achieve a deeper aggregation of information, while the latter refines input features through a feature transformation strategy. The progression of these models showcases the field's shift from basic fusion strategies to more complex attention mechanisms and sophisticated feature optimization.

4.2. The Deep Adaptation Challenge of Modal Heterogeneity Comparison of Experimental Results and Performance of Existing Models

The essential characteristics of text, visual, and audio differ significantly: text relies on discrete semantic symbols, such as the emotional word 'joy', visual relies on continuous spatiotemporal features, such as frame level changes in facial muscle movement, while audio is limited by dynamic fluctuations in acoustic features, such as the non-linear correlation between intonation and emotional intensity. This heterogeneity not only leads to temporal asynchrony (such as alignment deviation between video 30 frames per second and audio 16kHz sampling rate), but also causes semantic conflicts - for example, in the scenario of "neutral text+angry tone", traditional attention mechanisms are difficult to dynamically balance the weight of text semantics and audio emotional clues, which can easily lead to decision bias.

4.3. Scarcity and Generalization Bottleneck of Annotated Data

Multimodal emotion annotation requires cross modal expertise, such as psychological background annotators simultaneously interpreting text, expressions, and intonation, resulting in limited high-quality datasets. For example, IEMOCAP only contains 10 sets of dialogue data, and annotation consistency is significantly affected by subjective factors, such as differences in the definition of "surprise" and "fear". This directly leads to a sharp drop in the model's cross scene generalization performance. For example, when a model trained on the natural scene dataset Affwild2 is transferred to the professional actor dataset RAVDESS, the unweighted accuracy rate (UAR) decreases by an average of over 20%

4.4. Insufficient Interpretability of the Attention Mechanism

The existing visualization of attention weights, such as heatmaps, can only reflect the surface correlations of modal contributions and cannot reveal the deep logic of why the modality is given high weights. For example, the model may assign high visual weight to the "frown expression", but cannot explain the causal relationship between this expression and the "anger" emotion. This "black box" characteristic restricts its credibility in sensitive fields such as healthcare and education.

4.5. Prospect

In response to the above challenges, future research can make breakthroughs in three aspects:

One is to build a dynamic modal adaptation framework: combining meta learning to train a "modal credibility evaluator", which can judge the dominant modality of the scene in real time through a small number of samples, such as enhancing audio weights when there is text ambiguity, and introducing contrastive loss to optimize semantic alignment of heterogeneous features, alleviating temporal asynchrony and semantic conflicts.

The second is to explore lightweight data augmentation strategies: using generative models such as multimodal GANs to synthesize scarce samples, such as emotional expressions in niche cultures, combined with contrastive learning to learn modality invariant features, reducing dependence on annotated data, and improving cross scene generalization ability - for example, by generating "anger expressions+text" samples from different cultural backgrounds, enhancing the model's adaptability to the diversity of emotional expressions.

The third is to enhance the interpretability of attention mechanisms: introducing causal reasoning to trace the logical chain of decision-making (such as the correlation basis of "frowning expression → facial action unit → anger emotion"), developing interactive visualization tools to present key feature contributions in real time (such as highlighting the synergistic effect of "high tone+negative vocabulary"), making the decision-making process traceable and verifiable, and improving the feasibility of implementation in sensitive areas.

5. Conclusion

This review systematically reviews the research progress and development trends of attention mechanisms in the field of multimodal emotion recognition from 2017 to 2025. Research has shown that this field has gone through four distinct stages of evolution from basic feature fusion to multi-scale cross fusion, gradually establishing a technical system centered on attention refinement and modality interaction depth. By constructing a technical matrix of four types of attention mechanisms: cross modal, recursive, multi-scale, and dynamic, the study systematically elucidates the key roles of each mechanism in solving problems such as modal heterogeneity and feature alignment.

The current research still faces three main challenges: the problem of inconsistent feature distribution caused by essential differences between modalities has not been fully resolved; The scarcity of high-quality annotated data constrains the generalization performance of the model; The insufficient interpretability of attention weights limits their application in sensitive fields. These issues have become important bottlenecks hindering further technological development.

The innovative contribution of this article lies in proposing a dual dimensional classification framework and a four stage evolution model, providing a systematic perspective for understanding the development of attention mechanisms in multimodal emotion recognition. The constructed technology matrix not only clarifies the characteristics and applicable scenarios of various attention mechanisms, but also provides a theoretical basis and methodological reference for subsequent research.

Future research can focus on three directions: developing a dynamic modal adaptation framework based on meta learning to enhance the model's ability to handle heterogeneous data; Explore generative data augmentation methods to alleviate the scarcity of annotated data; Introduce causal reasoning mechanism to enhance the interpretability of attention weights and decision transparency. These research directions will help promote the development of multimodal emotion recognition technology towards more efficient, reliable, and practical directions, providing stronger technical support for application scenarios such as human-computer interaction and mental health monitoring

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Chen L, Ouyang Y, Zeng Y, Li Y. Dynamic facial expression recognition model based on BiLSTM-Attention. In: 2020 15th International Conference on Computer Science & Education (ICCSE). 2020: 828-832.

[2] Wu M, Su W, Chen L, Pedrycz W, Hirota K. Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition. IEEE Transactions on Affective Computing, 2020.

[3] Ameen S Y, Nourildean S W. Coordinator and router investigation in IEEE802.15.14 ZigBee wireless sensor network. In: 2013 International Conference on Electrical Communication, Computer, Power, and Control Engineering (ICECCPCE). 2013: 130-134.

[4] Al-Sultan M R, Ameen S Y, Abduallah W M. Real time implementation of Stegofirewall system. International Journal of Computing and Digital Systems, 2019, 8: 498-504.

[5] Zhang H. Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder. IEEE Access, 2020, 8: 164130-164143.

[6] Al-Naima F, Ameen S Y, Al-Saad A F. Destroying steganography content in image files. In: IEEE Proceedings of Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing. University of Patras, Patras, Greece, 2006.

[7] Cai C, He Y, Sun L, et al. Multimodal sentiment analysis based on recurrent neural network and multimodal attention. In: Proceedings of the 2nd Multimodal Sentiment Analysis Challenge. 2021: 61-67.

[8] Praveen R G, Alam J. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. arXiv preprint arXiv:2403.13659, 2024.

[9] G M H, S S S, V R S. Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN. International Journal of Interactive Multimedia and Artificial Intelligence, 2021, 6(6): 112-121.

[10] Luo H, Ji L, Huang Y, Wang B, Ji S, Li T. ScaleVLAD: Improving multimodal sentiment analysis via multi-scale fusion of local descriptors. arXiv preprint arXiv:2112.01368, 2021.

[11] Khan M, Tran P N, Pham N T, El Saddik A, Othmani A. MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion. Scientific Reports, 2025, 15: 5473.

[12] Zhao J, Wei X, Bo L. R1-Omni: Explainable omni-multimodal emotion recognition with reinforcement learning. arXiv preprint arXiv:2503.05379, 2025.

[13] Feng X, Lin Y, He L, et al. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. arXiv preprint arXiv:2410.04491, 2024.

[14] Yang L, Zhong J, Wen T, et al. CCIN-SA: Composite cross modal interaction network with attention enhancement for multimodal sentiment analysis. Information Fusion, 2025, 123: 103230.