Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.076411

REVIEW

A Survey on Multimodal Emotion Recognition: Methods, Datasets, and Future Directions

A-Seong Moon, Haesung Kim, Ye-Chan Park and Jaesung Lee*

Department of Artificial Intelligence, Chung-Ang University, Seoul, Republic of Korea

*Corresponding Author: Jaesung Lee. Email: curseor@cau.ac.kr

Received: 20 November 2025; Accepted: 19 January 2026; Published: 12 March 2026

ABSTRACT: Multimodal emotion recognition has emerged as a key research area for enabling human-centered artificial intelligence, supported by the rapid progress in vision, audio, language, and physiological modeling. Existing approaches integrate heterogeneous affective cues through diverse embedding strategies and fusion mechanisms, yet the field remains fragmented due to differences in feature alignment, temporal synchronization, modality reliability, and robustness to noise or missing inputs. This survey provides a comprehensive analysis of MER research from 2021 to 2025, consolidating advances in modality-specific representation learning, cross-modal feature construction, and early, late, and hybrid fusion paradigms. We systematically review visual, acoustic, textual, and sensor-based embeddings, highlighting how pre-trained encoders, self-supervised learning, and large language models have reshaped the representational foundations of MER. We further categorize fusion strategies by interaction depth and architectural design, examining how attention mechanisms, cross-modal transformers, adaptive gating, and multimodal large language models redefine the integration of affective signals. Finally, we summarize major benchmark datasets and evaluation metrics and discuss emerging challenges related to scalability, generalization, and interpretability. This survey aims to provide a unified perspective on multimodal fusion for emotion recognition and to guide future research toward more coherent and generalizable multimodal affective intelligence.

KEYWORDS: Multimodal emotion recognition; multimodal learning; cross-modal learning; fusion strategies; representation learning

1 Introduction

Emotions serve as essential elements in shaping human cognition and behavior. These affective states influence perception, guide decision-making, and govern both verbal and non-verbal communication. As intelligent systems continue to advance and permeate daily life, the capacity to recognize and interpret human emotions has become a vital component of human-centered artificial intelligence [1]. Emotion recognition enables machines to engage with users in a more adaptive and context-aware manner, fostering natural interaction across domains such as healthcare [2], education [3], and autonomous systems [4]. Over the past decade, automatic emotion recognition has evolved from unimodal frameworks that relied solely on audio or visual cues to multimodal systems that integrate diverse affective signals to capture the complexity of human emotion [5,6]. This transition has positioned fusion strategies at the core of emotion recognition research, shaping how information from multiple modalities is processed, aligned, and interpreted within computational models.

The integration of multiple modalities within emotion recognition systems has been motivated by the complementary nature of affective cues. Visual signals convey facial expressions and gestures, while acoustic signals capture prosodic and tonal variations that reflect subtle emotional nuances. When combined, these modalities provide a richer and more reliable understanding of human affect than either source alone. However, the process of merging heterogeneous data streams raises a fundamental design question: how and when should information from different modalities be fused? Over the years, research has converged on three principal paradigms: early fusion, late fusion, and hybrid fusion [7]. Early fusion aggregates features before inference, enabling joint learning of cross-modal relationships at the representation level. Late fusion, in contrast, combines independently learned modality-specific predictions, emphasizing interpretability and modularity. Hybrid fusion seeks to balance these two approaches by introducing intermediate-level interactions through mechanisms such as attention, gating, or transformer-based alignment. Each paradigm embodies a distinct trade-off among expressiveness, flexibility, and computational efficiency, forming the foundation for most MER frameworks.

In emotion recognition research, an equally important consideration lies in how emotions are represented. Broadly, affective states are modeled using either categorical or dimensional formulations. Categorical emotion models describe emotions as discrete classes, such as happiness, sadness, anger, or fear, and have been widely adopted in early emotion recognition systems due to their intuitive interpretability. In contrast, dimensional emotion models characterize affect along continuous axes, most commonly valence and arousal, where valence reflects the degree of emotional positivity or negativity and arousal indicates the level of activation or intensity. This representation enables finer-grained modeling of emotional dynamics and ambiguity, and is particularly well suited to multimodal and conversational settings where emotions evolve gradually over time. As a result, both categorical and dimensional formulations coexist in modern MER research, with the choice of representation influencing dataset design, learning objectives, and fusion strategies.

To contextualize these fusion paradigms within the broader MER workflow, Fig. 1 presents a standard processing pipeline commonly employed in multimodal systems. The pipeline processes raw inputs from visual, audio, text, and sensor modalities and derives informative representations from each stream. These representations are subsequently integrated through an appropriate fusion strategy, either early, late, or hybrid, before the consolidated embedding is mapped to the final emotion prediction space.

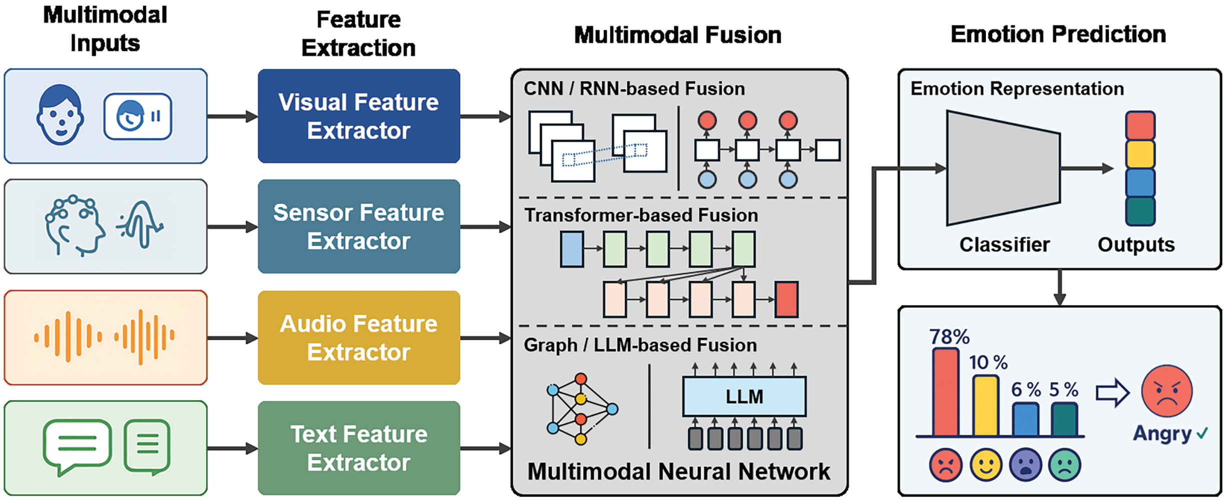

Figure 1: Overview of the multimodal emotion recognition (MER) pipeline.

Although fusion has become the cornerstone of MER, designing an effective integration mechanism remains a persistent challenge. Early fusion approaches enable joint representation learning but often struggle with the heterogeneity of feature distributions and the dominance of certain modalities during training. Late fusion methods, while robust to such discrepancies, tend to overlook fine-grained temporal dependencies that are essential for interpreting dynamic emotional expressions. These limitations have led researchers to explore hybrid fusion strategies that incorporate intermediate interactions between modalities. By leveraging mechanisms such as cross-modal attention, adaptive weighting, and temporal alignment, hybrid fusion frameworks attempt to capture both local and global dependencies across modalities. The continuous evolution of these methods reflects an ongoing effort to balance interpretability, robustness, and adaptability—key factors that determine the overall effectiveness of MER systems.

Given the diversity of fusion strategies and their pivotal role in the shape of MER, a systematic examination of these approaches is essential to understand the current research landscape. This survey aims to provide a comprehensive analysis of fusion methodologies, categorizing existing studies by integration stage, architectural design, and interaction dynamics among modalities. In doing so, it highlights how different fusion paradigms influence the learning of emotional representations and the interpretability of affective responses. Furthermore, this paper discusses emerging trends such as attention-based fusion, transformer-driven architectures, and adaptive gating mechanisms that redefine the boundaries between early, late, and hybrid designs. By consolidating the existing literature and identifying conceptual links between studies, this survey aims to offer a unified perspective on multimodal fusion for emotion recognition and to guide future research toward more cohesive, generalizable integration frameworks.

2 Related Work

Recent studies in MER have expanded rapidly, encompassing developments in architecture design, feature representation, and data-driven evaluation. This section reviews these advancements across four complementary perspectives. Section 2.1 introduces recent trends in MER, outlining how deep learning frameworks and multimodal architectures have evolved in recent years. Section 2.2 discusses embedding and feature representation methods, focusing on how pre-trained encoders and shared latent spaces enhance the expressiveness of affective cues across modalities. Section 2.3 examines various fusion strategies, comparing early, late, and hybrid approaches and analyzing how each contributes to the integration of multimodal information. Finally, Section 2.4 summarizes the datasets and evaluation metrics commonly used in the field, highlighting their roles in benchmarking performance and facilitating consistent comparison across studies.

2.1 Recent Advances in MER

To provide a systematic overview of recent progress in MER, representative studies published between 2021 and 2025 are organized and analyzed with respect to their core modalities, fusion strategies, pre-trained feature extractors, and benchmark datasets. A comprehensive summary of these works is provided in Table A1 in Appendix A for reference. In this subsection, the major works are discussed in both thematic and temporal order to highlight their technical motivations, fusion mechanisms, and contributions to multimodal understanding, while emphasizing the methodological shifts from early handcrafted fusion designs to transformer- and large-language-model-based architectures.

The studies included in this survey were selected through a structured literature review process. We primarily targeted peer-reviewed journals and top-tier conference publications in the fields of affective computing, multimodal learning, and human-centered AI. A literature search was conducted across common academic databases using keywords such as multimodal emotion recognition, emotion recognition in conversation, audio-visual-text fusion, and multimodal affective computing. Priority was given to works that introduced novel fusion strategies, alignment mechanisms, or architectural contributions, as well as studies evaluated on widely used public benchmarks. This selection strategy resulted in a curated set of representative papers that collectively reflect recent methodological trends in MER.

Early multimodal approaches established hybrid fusion as the foundation of modern MER. Progressive and cross-modal reinforcement frameworks improved unaligned sequence modeling through attention- and message-based interactions, improving the robustness of emotion inference under modality asynchrony [8,9]. Hybrid integration of convolutional and recurrent encoders enabled richer temporal reasoning across speech, text, and facial modalities, achieving balanced feature complementarity [10,11]. In 2022, research shifted toward disentangled and uncertainty-aware architectures. Feature-disentangled multimodal learning separated shared and modality-specific components to mitigate redundancy and distribution gaps, while hierarchical uncertainty modeling and self-supervised pretraining improved robustness under data imbalance [12–14]. Meanwhile, unsupervised and sensor-integrated systems extended MER beyond audiovisual signals by incorporating physiological cues and radar sensing, achieving strong generalization across real-world settings [15–17]. Collectively, these developments marked the transition from handcrafted concatenation to semantic-aware fusion and modality disentanglement.

The 2023 research phase marked the consolidation of transformer-based architectures and conversational emotion recognition. Hybrid transformers such as DF-ERC, MER-HAN, and SDT modeled intra- and inter-modal dependencies through attention, self-distillation, and context-aware gating, achieving state-of-the-art accuracy on MELD and IEMOCAP datasets [18–20]. Multi-label and emotion-level embedding frameworks introduced emotion co-occurrence modeling and simultaneous classification of multiple affective states [21,22]. Parallel trends focused on label efficiency and robustness under missing modalities. Semi-supervised and uncertainty-calibrated methods such as Expression-MAE, COLD Fusion, and IF-MMIN leveraged pseudo-labeling and modality imagination to sustain performance with limited supervision [23–25]. Complementary research emphasized audiovisual synchronization and deep metric learning through cross-modal attention and Bi-GRU modeling, bridging acoustic–visual coherence in low-resource environments [26–28]. Altogether, 2023 consolidated transformer-based fusion as a dominant paradigm, emphasizing interpretability, resilience, and fine-grained interaction across modalities.

The 2024 research landscape represented a pivotal shift toward instruction-tuned and graph-augmented architectures integrating large language models (LLMs) and cross-modal reasoning. LLM-based textual encoders have recently advanced MER toward reasoning-level affect understanding by enabling richer semantic abstraction and contextual modeling. Through instruction-tuned and dialogue-aware pretraining, these models can capture discourse coherence, implicit emotional cues, and causal relationships that extend beyond surface-level sentiment expressions. Such capabilities are particularly beneficial in conversational MER scenarios, where emotional states are influenced by long-range context, speaker intent, and pragmatic nuances. As a result, LLM-based frameworks provide a powerful mechanism for modeling complex emotional dynamics that are difficult to capture with conventional sequence encoders. Instruction-tuned systems such as Emotion-LLaMA and DialogueMLLM unified audio, visual, and textual modalities through structured prompting and generative reasoning [29,30]. Graph-based designs such as MGLRA, AGF-IB, and DEDNet enhanced contextual alignment and speaker dependency modeling in dialogues [31–33]. Hybrid interpretable architectures, including CNN–BERT fusion and token-disentangling transformers, increased transparency and localization while maintaining high accuracy [34,35]. Interpretable systems like ParallelNet and KoHMT extended multimodal fusion into cross-lingual and domain-specific contexts [36,37], while sparsity- and uncertainty-based networks dynamically filtered redundant cues to improve robustness [24,38]. Domain-oriented research expanded MER into healthcare and physiological analysis by incorporating electroencephalography (EEG) and remote photoplethysmography (rPPG)—based fusion, supported further by cross-subject generalization networks [2,39,40]. Calibration and transfer-learning methods such as CDaT and CMERC refined confidence estimation and semantic consistency across conversational datasets [41,42].

The representational benefits of LLM-based MER come at substantial computational cost. Large parameter scales, deep transformer stacks, and multi-stage inference pipelines increase memory consumption, energy usage, and inference latency, which limits applicability in real-time, on-device, or large-scale deployment scenarios. In contrast, lighter hybrid fusion architectures that combine compact textual encoders with attention- or gating-based multimodal integration mechanisms often achieve a more favorable balance between expressive capacity and computational efficiency. This contrast underscores a fundamental design trade-off in contemporary MER, where the depth of reasoning enabled by LLMs must be carefully weighed against efficiency, scalability, and deployment constraints. Collectively, these considerations highlight that 2024 research consolidated MER not only through LLM integration and graph reasoning, but also by clarifying the practical boundaries between expressive power and computational feasibility.

The 2025 research wave continued the evolution of MER toward foundation-model–oriented architectures emphasizing generative recovery, unsupervised learning, and explainability. Transformer-driven frameworks such as MemoCMT and RMER-DT leveraged cross-modal transformers and diffusion-based restoration to reconstruct missing modalities and capture long-range conversational dependencies [43,44]. Graph-spectrum and motion-aware methods, such as GS-MCC and MIST, refined relational and temporal fusion across multimodal signals [45,46]. Explainable and lightweight systems adopted Gradient-SHAP–based feature attribution and EEG–facial fusion to enhance transparency and efficiency without sacrificing performance [47–49]. Weakly supervised frameworks such as MGAFR and MERITS-L combined graph aggregation, contrastive learning, and LLM-guided pretraining for robust adaptation across incomplete datasets [50,51]. Meanwhile, spiking-transformer hybrids such as SPSNCVT introduced bio-inspired temporal encoding to defend against adversarial perturbations [52], while multi-granularity and mixture-of-experts models advanced fine-grained alignment and dynamic expert gating [53–55]. Collectively, 2025 research established MER as a convergence of foundation-model fusion, generative recovery, and explainable reasoning, signaling its transition toward scalable and human-aligned affective intelligence.

To provide a consolidated view of recent progress, Table 1 summarizes the performance results reported from representative MER studies published in the last five years. The results are organized by evaluation metric, with datasets presented side by side within each table to facilitate comparison across commonly used benchmarks. For each dataset, a single evaluation metric is selected based on prevalent reporting practices in the literature, and each row corresponds to an individual study. Only dataset–metric pairs reported by at least ten papers are included to ensure that the comparison reflects sufficiently established empirical trends. All values are reproduced from the original publications under their respective experimental settings and are intended to illustrate performance tendencies rather than establish a unified benchmark.

Table 1: Reported performance snapshots on public MER datasets. Values are reproduced from the original studies under their respective experimental settings.

Dataset	Metric	Year	Fusion Approach	Score	Ref.
CMU-MOSEI	F1-score	2025	Hybrid Fusion	86.17	[51]
		2024	Early Fusion	87.90	[34]
		2024	Hybrid Fusion	86.50	[35]
		2024	Late Fusion	83.00	[56]
		2023	Hybrid Fusion	46.10	[24]
		2022	Hybrid Fusion	85.80	[12]

(Continued)

Table 1 (continued)

Dataset	Metric	Year	Fusion Approach	Score	Ref.
		2022	Hybrid Fusion	75.90	[17]
		2022	Hybrid Fusion	69.50	[57]
		2021	Hybrid Fusion	86.23	[10]
		2021	Hybrid Fusion	82.60	[8]
MELD	F1-score	2025	Early Fusion	69.00	[45]
		2025	Hybrid Fusion	67.02	[44]
		2025	Late Fusion	66.02	[50]
		2024	Hybrid Fusion	67.02	[58]
		2024	Early Fusion	66.85	[42]
		2024	Hybrid Fusion	65.76	[33]
		2023	Hybrid Fusion	66.60	[20]
		2023	Hybrid Fusion	60.22	[19]
		2022	Hybrid Fusion	58.56	[13]
		2021	Late Fusion	64.00	[9]
IEMOCAP	Accuracy	2025	Hybrid Fusion	81.33	[43]
		2025	Hybrid Fusion	80.24	[53]
		2024	Hybrid Fusion	71.72	[58]
		2023	Early Fusion	85.90	[21]
		2023	Hybrid Fusion	82.70	[24]
		2023	Late Fusion	82.57	[59]
		2023	Hybrid Fusion	79.71	[60]
		2023	Hybrid Fusion	77.00	[61]
		2022	Hybrid Fusion	69.60	[62]
		2021	Hybrid Fusion	79.77	[63]
		2021	Hybrid Fusion	61.80	[11]
RAVDESS	Accuracy	2025	Hybrid Fusion	88.10	[64]
		2024	Hybrid Fusion	95.00	[65]
		2024	Early Fusion	93.18	[66]
		2024	Late Fusion	81.90	[67]
		2023	Early Fusion	93.23	[26]
		2023	Hybrid Fusion	89.25	[27]
		2022	Late Fusion	94.99	[68]
		2022	Hybrid Fusion	93.17	[57]
		2022	Late Fusion	86.00	[69]
		2021	Late Fusion	86.70	[70]
		2021	Hybrid Fusion	80.08	[71]

2.2 Embedding and Feature Representation

Emotion recognition relies heavily on how features are extracted and represented across heterogeneous modalities. Effective embedding enables the model to capture both low-level perceptual signals and high-level semantic patterns that reflect emotional states. This section categorizes feature representation methods into five perspectives, visual, audio, textual, sensor-based, and cross-modal, highlighting how each modality contributes to a comprehensive affective understanding and how feature integration before fusion enhances multimodal robustness and interpretability.

Visual Feature Embedding

Visual feature embedding serves as the primary channel for interpreting facial and bodily expressions that reflect subtle emotional states in MER. Early visual frameworks relied heavily on convolutional neural networks such as VGG and ResNet to extract spatial features from facial images, providing robust representations of muscle movements and expression intensities [72,73]. These methods primarily focused on static cues, achieving stable results in controlled environments but showing limited adaptability in dynamic, spontaneous contexts.

To capture temporal information, later architectures extended spatial encoders with recurrent or 3D convolutional layers. For instance, hybrid CNN–BiLSTM and 3D-CNN models enabled spatio-temporal learning from sequential frames, successfully modeling emotion transitions in datasets such as RAVDESS and CK+ [46,62]. This evolution reflected the growing awareness that emotion perception is inherently dynamic, requiring models to track subtle temporal variations in facial action units. Some approaches also leveraged optical flow and motion vectors to supplement static frame representations, enriching temporal consistency in recognition outcomes [65].

Attention-based visual modules subsequently emerged to emphasize discriminative regions and suppress irrelevant variations caused by lighting or pose. Spatial and channel attention frameworks selectively amplified emotion-relevant facial areas, while transformer-based mechanisms learned context-aware global dependencies [27,74]. These designs significantly improved feature interpretability and generalization, particularly in in-the-wild datasets characterized by occlusion and heterogeneous subjects.

The introduction of Vision Transformers (ViT) marked a further paradigm shift in visual feature extraction. ViT and its hybrid variants captured long-range dependencies across facial regions, while capsule and graph-transformer hybrids improved relational reasoning by modeling inter-feature connectivity [75,76]. Such methods demonstrated higher robustness to complex emotional scenes, especially those involving microexpressions, head movements, and multi-person interactions.

Recently, visual embeddings have become tightly integrated into unified frameworks. Emotion-LLaMA [29] leveraged instruction-tuned learning to align visual, acoustic, and textual inputs for emotion reasoning, while MIST [46] combined ResNet-50 and 3D-CNN modules to analyze facial appearance and motion cues jointly. Other studies have sought to enhance data efficiency through self-supervised and cross-domain pretraining, incorporating contrastive or generative objectives to address limited labeled data and domain variability [48,73]. Collectively, these advancements reveal that visual embedding research has evolved from static CNN representations toward transformer-based, semantically aligned, and pretraining-enhanced paradigms that support scalable multimodal affect understanding.

Audio Feature Embedding

Audio feature embedding captures prosodic, spectral, and paralinguistic cues that reflect affective states such as tone, rhythm, and vocal intensity. Early studies primarily relied on handcrafted acoustic features, including Mel-Frequency Cepstral Coefficients, pitch, energy, and zero-crossing rate, which were processed through classical classifiers or shallow neural networks [68,72]. Although these features provided interpretable descriptors of emotion-related variations, they lacked the ability to represent high-level semantics or contextual dependencies within speech.

The advent of deep learning expanded the expressiveness of acoustic representations through convolutional and recurrent encoders. CNN-based architectures were utilized to learn discriminative spectro-temporal patterns directly from Mel-spectrograms, capturing fine-grained frequency transitions and amplitude dynamics [2,65]. These models improved generalization across diverse speakers and recording environments. To capture long-range dependencies and rhythm-sensitive structures, BiLSTM and GRU layers were often appended to CNN backbones, effectively modeling temporal continuity in emotional speech sequences [39,62].

Self-supervised and transformer-based encoders have since redefined the audio embedding paradigm. Pretrained speech models such as wav2vec2.0, HuBERT, and PANNs emerged as dominant backbones due to their capacity to represent phonetic and prosodic variations without extensive labeled data [29,37,49]. These models enabled more transferable emotion features by capturing latent patterns in pitch contour and energy modulation across multiple corpora. Studies such as MemoCMT [43] and MGCMA [53] further demonstrated that transformer-based acoustic embeddings preserve cross-temporal coherence, allowing stable alignment with linguistic content and contextual affect transitions.

Attention mechanisms have also played a crucial role in refining the representation of acoustic emotions. Channel and time-domain attention modules dynamically reweighted salient frequency bands to enhance emotion sensitivity while filtering noise and neutral speech segments [32,34]. More recent designs integrated multi-scale or token-level attention to focus on rhythm discontinuities, emotional bursts, and silence intervals that correlate with affective intensity [35,55]. These methods bridged low-level spectral variation with higher-level emotional intent, significantly improving discrimination among subtle emotions such as anxiety, surprise, and contempt.

Recent research has also explored bioacoustic and physiological speech correlates to expand the affective dimension of auditory representation. Studies combining acoustic features with auxiliary sensor data, such as respiratory or vocal muscle signals, have shown enhanced robustness under noisy or spontaneous conditions [15,16]. Furthermore, end-to-end audio pipelines employing variational or diffusion-based augmentation strategies have emerged to increase resilience against data imbalance and adversarial perturbations [52,77].

Taken together, advancements in audio embedding have transitioned from handcrafted spectral analysis to deep transformer-driven architectures that model emotion as a structured, temporally evolving process. The evolution of pretrained speech encoders and attention-based refinement has made acoustic representation a crucial foundation for capturing implicit affective dynamics in MER systems.

Text Feature Embedding

Textual feature embedding provides the semantic and syntactic foundation for understanding the affective meaning conveyed by linguistic expressions, such as word choice, sentence structure, and discourse context. In MER, text serves as a high-level cue that complements perceptual modalities by explicitly expressing emotions through sentiment, appraisal, or figurative language. Early approaches employed word-level features such as bag-of-words and TF–IDF vectors or used static distributed representations like Word2Vec and GloVe to encode emotional semantics within utterances [11,62]. These representations enabled initial progress in textual sentiment classification but lacked context awareness and failed to capture polarity shifts across sentences.

The introduction of deep neural encoders such as CNNs, RNNs, and Bi-LSTMs enhanced the capacity to model sequential dependencies and contextual nuances in emotional text [18,34]. Hierarchical recurrent networks further incorporated dialogue-level context, enabling emotion understanding in conversations where prior utterances influence the affective meaning of later ones. Despite these advancements, such models remained constrained by their reliance on fixed context windows and struggled to represent long-range semantic dependencies or subtle pragmatic cues such as sarcasm and irony.

Transformer-based encoders have fundamentally reshaped textual emotion representation by introducing self-attention mechanisms that capture global dependencies and polysemous word usage. As LLMs proliferate, pretrained language models such as BERT, RoBERTa, DeBERTa, and ELECTRA have become dominant backbones for textual embedding in MER, offering fine-grained contextual representations of emotional polarity and intensity [37,58,60]. These models learn not only lexical associations but also discourse-level sentiment trajectories, allowing them to align with temporal or conversational structures when combined with other modalities. Moreover, fine-tuning on emotion-specific corpora such as IEMOCAP, MELD, and CMU-MOSEI has proven effective for adapting general-purpose transformers to affective language tasks.

Beyond single-stream textual encoding, recent research emphasizes semantic disentanglement and cross-modal alignment at the embedding level. Models such as FDRL [78] and CAMEL [75] decouple shared and modality-specific semantic spaces, enabling robust representation of emotional meaning even in metaphorical or abstract expressions. Other approaches integrate linguistic prosody by aligning textual embeddings with the corresponding acoustic cues, forming joint representations aware of emotions that account for explicit and implicit affect [38,43].

LLMs–based architectures have recently extended textual embedding toward reasoning-level affect understanding. Systems such as Emotion-LLaMA [29] and DialogueMLLM [30] incorporate instruction-tuned and dialogue-aware pretraining to interpret emotional intent, causal relations, and contextual shifts. These frameworks highlight an emerging trend in textual embedding—transitioning from surface-level sentiment extraction to reasoning-driven affect interpretation that supports generalized multimodal emotion understanding.

Sensor Feature Embedding

Sensor-based feature embedding focuses on physiological and behavioral signals that reflect internal affective states beyond visible or auditory expression. Typical modalities include EEG, electrocardiography (ECG), galvanic skin response (GSR), electromyography (EMG), motion capture, and rPPG. These biosignals capture autonomic activity patterns related to arousal, stress, and emotion, enabling the interpretation of latent affective cues that are not directly observable.

Early approaches relied on convolutional and recurrent architectures to extract time–frequency and temporal dynamics from multi-channel signals. For instance, depthwise separable CNNs and Bi-LSTM models were used to model spatial and temporal dependencies in EEG and ECG data, improving recognition accuracy in healthcare-related affective analytics [2]. Recent architectures such as graph neural networks and transformers have been employed to model spatial correlations among electrodes or motion nodes, as demonstrated in multimodal physiological frameworks like MEmoR [16] and TDTN-HLFR [79], which integrate physiological and visual cues for robust emotion inference.

Despite their effectiveness, sensor-based embeddings face challenges due to signal noise, inter-subject variability, and the limited availability of large-scale labeled datasets. To address these issues, recent studies have adopted normalization, domain adaptation, and adversarial training strategies to enhance cross-subject generalization. For example, the CAG-MoE framework [54] integrates multiple sensor streams using cross-attention and gated mixtures of experts to adaptively weight modalities in the presence of noisy or missing data. These advances highlight a shift toward multimodal physiological fusion, combining sensors with visual or acoustic cues to achieve interpretable and context-aware emotion recognition across diverse environments.

Cross-Modal Feature Representation

Cross-modal feature representation aims to align heterogeneous modality embeddings into a unified latent space where shared affective semantics can be effectively captured. Unlike unimodal encoders that learn features independently, this approach focuses on discovering common structures and correlations between modalities such as audio, visual, text, and physiological signals. By projecting them into a shared embedding space, models can transfer emotion-relevant information across modalities, enabling robust recognition even under degraded or missing conditions.

Early studies used deep canonical correlation analysis and autoencoder-based projection to align multimodal features, as seen in MERDCCA [11] and MCWSA-CMHA [80], which enforced correlation consistency between GRU- or CNN-encoded features. More recent frameworks introduced contrastive and self-supervised objectives, exemplified by modality-pairwise contrastive learning [57] and contrastive modality-invariant representation models like IF-MMIN [25] and CIF-MMIN [81], which improved robustness to missing or corrupted modalities. Other approaches, such as CAMEL [75], extended this idea by incorporating metaphor-aware contrastive alignment and disentangled contextual learning to capture higher-order semantic relationships between modalities.

In parallel, the rise of large pre-trained encoders and multimodal foundation models has accelerated the use of CLIP-style alignment and prompt-tuned joint embeddings. Architectures such as MEmoBERT [14], KoHMT [37], and DialogueMLLM [30] demonstrate how instruction-tuned or cross-modal transformer layers unify feature spaces across modalities, enabling emotion recognition and reasoning through shared attention mechanisms. These models often leverage knowledge distillation or adversarial feature alignment to ensure modality consistency while preserving discriminative affective cues.

Recent research trends increasingly emphasize the transition from simple feature concatenation toward semantically aligned representation learning, where emotion-relevant patterns are disentangled, normalized, and jointly optimized across modalities. Although computationally intensive, such alignment-driven methods serve as a conceptual bridge between feature-level fusion and foundation-level multimodal reasoning, marking an essential step toward scalable, generalizable emotion-understanding systems.

2.3 Fusion Approaches in MER

Fusion serves as the core mechanism through which MER systems integrate heterogeneous affective cues into a unified representation space. As summarized in Fig. 2, existing studies can be broadly grouped into three paradigms: early fusion, late fusion, and hybrid fusion. Early fusion aggregates features before inference; late fusion combines modality-specific decisions at the score level; and hybrid fusion introduces intermediate interactions via attention, graph, or transformer-based modules.

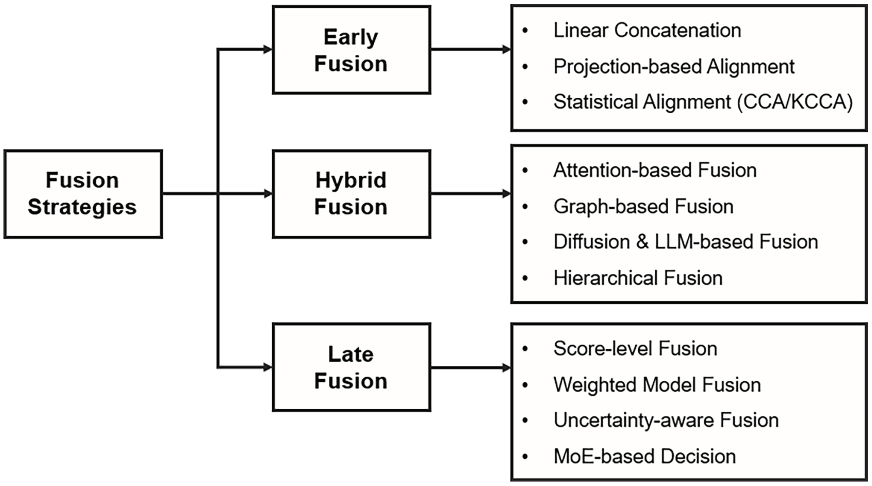

Figure 2: Taxonomy of fusion strategies in MER, organized into early, late, and hybrid fusion with representative implementation patterns.

Fusion serves as the central mechanism through which MER systems integrate heterogeneous affective cues into coherent representations. As discussed in Section 1, the effectiveness of these systems largely depends on how and when modalities are combined. Existing research has converged on three dominant paradigms—early, late, and hybrid fusion—each reflecting a distinct design philosophy and trade-off between representational richness and interpretability. Early fusion strategies concatenate or jointly encode raw or low-level features prior to inference, promoting deep cross-modal interactions but facing challenges from distributional mismatches. Late fusion aggregates modality-specific predictions at the decision level, allowing modular training and interpretability but often neglecting temporal and contextual dependencies. Hybrid fusion introduces intermediate-level integration through attention, gating, or transformer-based alignment mechanisms, balancing modality specialization with joint representation learning.

Before 2021, MER studies predominantly relied on early and late fusion schemes that combined handcrafted acoustic and visual features or aggregated classifier outputs at the decision level. As deep neural networks and transformer-based mechanisms matured, the field experienced a paradigm shift toward hybrid fusion, enabling adaptive and context-aware integration among modalities. Fig. 3 presents the distribution of fusion approaches adopted between 2021 and 2025. While early and late fusion remain relevant for interpretability and lightweight deployment, hybrid fusion dominates recent research, reflecting its capacity to model complex inter-modality relationships and temporal dependencies.

Early Fusion Approach

Early Fusion is the most direct strategy for integrating multimodal inputs, where heterogeneous affective cues such as audio, visual, and textual features are combined before inference. In this approach, the model receives concatenated or jointly encoded representations of multiple modalities, enabling it to learn cross-modal relationships within a unified embedding space. This joint optimization allows fine-grained emotional correlations, such as prosodic variations aligning with subtle facial expressions, to be captured at an early stage of learning. Despite these strengths, the design increases the risk of imbalance between modalities due to differences in feature distributions or temporal resolutions.

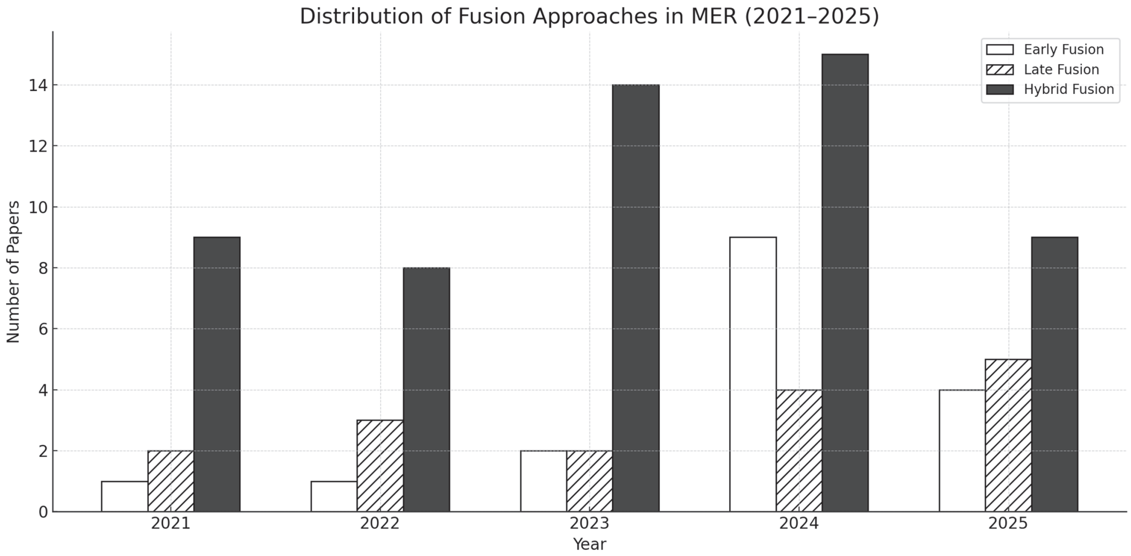

Figure 3: Trends in the adoption of early, late, and hybrid fusion approaches in MER from 2021 to 2025. Hybrid fusion has consistently maintained the highest adoption rate, signifying its role as the mainstream integration strategy in modern MER architectures.

In MER, early fusion played an essential role in the initial adoption of deep learning–based frameworks. Many early studies focused on combining acoustic and visual features through convolutional or recurrent architectures to capture both spatial and temporal emotional dynamics. For instance, CNN–BiLSTM pipelines effectively modeled frame-level facial features and spectral variations in speech, achieving synchronized recognition of expressive cues across modalities [26,65]. Later studies extended this concept to physiological signals such as EEG and rPPG, demonstrating that fusing visual and biosignal features enables more stable, noise-resistant affect prediction, particularly in healthcare and real-time monitoring environments [2,39,73]. Early fusion has also been applied to text–speech pairs, where embeddings from pretrained models such as HuBERT, BERT, or fastText are jointly processed to enhance efficiency in conversational emotion recognition tasks [29,56].

Early fusion can be summarized as a process in which modality-specific features are first extracted, aligned, and concatenated into a joint feature space before being passed to shared layers for classification. Fig. 4 illustrates this mechanism, showing how feature-level integration encourages direct cross-modal interaction but requires precise temporal alignment among inputs. Such direct concatenation supports dense inter-modality learning but often amplifies noise or redundancy when one modality dominates the shared representation. To address these issues, recent models incorporate lightweight normalization or attention-based balancing layers, aiming to retain the simplicity of early fusion while improving its stability in large-scale multimodal settings [35].

Overall, early fusion remains a foundational yet evolving paradigm in MER. Its strength lies in its ability to promote deep joint representation learning and real-time processing efficiency, making it particularly suitable for synchronized or resource-limited scenarios. Nevertheless, its sensitivity to misaligned or missing data and limited flexibility in dynamic weighting modalities have driven the emergence of hybrid fusion architectures, which integrate the representational depth of early fusion with adaptive mechanisms for modality control.

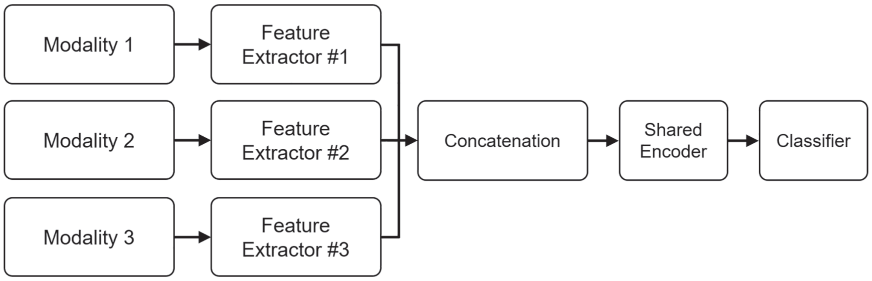

Figure 4: Early fusion integrates modality-specific features, such as audio, visual, and text, into a shared representation space prior to inference. This joint feature-level combination enables cross-modal learning but increases sensitivity to temporal misalignment and imbalance across modalities.

Late Fusion Approach

Late fusion refers to the integration of modality-specific predictions at the decision stage rather than during feature representation learning. In this paradigm, each modality, such as audio, visual, or text, is processed independently through its own encoder or classifier, and the resulting emotion probabilities are aggregated through ensemble mechanisms such as weighted averaging, majority voting, or trainable gating networks. This modular structure allows models to flexibly combine heterogeneous input sources and maintain interpretability, making late fusion particularly useful in scenarios where each modality performs reliably in isolation.

Representative implementations have demonstrated its practicality across diverse emotion recognition settings. For instance, ensemble-based audiovisual systems such as MIST [46] integrate ResNet-50, 3D-CNN, and Semi-CNN features through weighted decision aggregation, while real-time frameworks like Dixit et al. [56] employ 1D and 2D CNNs with text embeddings to achieve efficient inference through random forest fusion. Similarly, multimodal learning schemes such as Radoi et al. [67] improve robustness under resource constraints by combining CNN outputs that are aware of uncertainty from audio and visual streams.

As illustrated in Fig. 5, late fusion operates by first generating independent modality predictions that are subsequently merged at the decision level. Each stream outputs a distinct emotional probability distribution, and these are unified through ensemble or gating mechanisms to produce the final classification result. This structure enables flexible model substitution and modular retraining without altering the entire pipeline, a key advantage in large-scale or continually evolving affective computing systems.

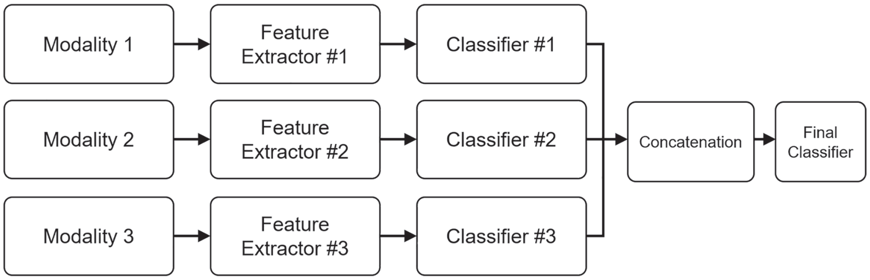

Figure 5: Late fusion combines modality-specific predictions, such as those from audio, visual, and text classifiers, at the decision level through weighted or ensemble aggregation. This design enhances modularity and interpretability but limits the ability to model fine-grained temporal and contextual dependencies across modalities.

Although late fusion offers modularity and flexibility, the separation between representation learning and decision integration limits the model's ability to capture temporal synchronization and semantic dependencies across modalities. Consequently, late fusion frameworks may struggle with tasks that require subtle context inference or emotional transition tracking. Recent approaches, such as CAG-MoE [54], attempt to address these limitations by incorporating cross-attention and expert gating mechanisms, bridging decision-level integration with intermediate representation learning.

Hybrid Fusion Approach

Hybrid fusion represents an intermediate paradigm that bridges early and late fusion by enabling cross-modal interactions at multiple stages of representation. Rather than fusing only raw features or final predictions, hybrid frameworks incorporate adaptive mechanisms that dynamically determine how modalities interact at the feature, intermediate, or decision level. As shown in Fig. 6, this paradigm integrates attention, gating, and transformer-based alignment modules to facilitate both modality-specific specialization and shared representation learning. Such flexibility allows hybrid fusion to balance the expressiveness of early fusion with the modular interpretability of late fusion, becoming the dominant design trend in MER.

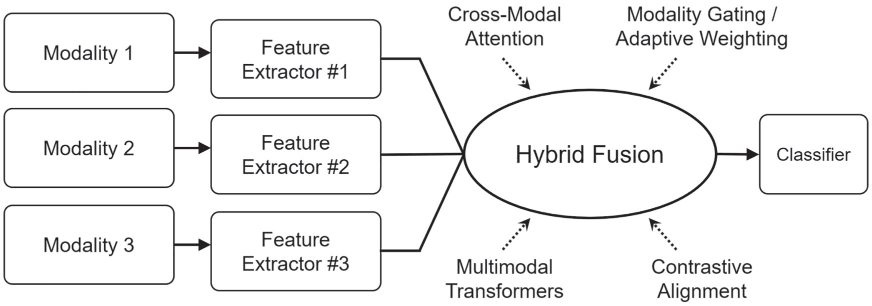

Figure 6: Hybrid fusion introduces intermediate-level interactions between modalities through attention, gating, or transformer-based alignment. This approach balances the complementary strengths of early and late fusion, capturing both fine-grained cross-modal dependencies and high-level semantic coherence.

Attention-based hybrid fusion utilizes selective weighting to highlight the most informative modality features while suppressing noisy or irrelevant signals. Cross-modal attention modules compute intermodal relevance, allowing one modality (e.g., audio) to conditionally refine another (e.g., visual) based on emotional salience. Early implementations employed hierarchical co-attention [19,20], where intra-modal attention first captured local dependencies before inter-modal attention fused complementary cues. More advanced variants, such as DF-ERC [18] and MER-HAN [19], integrated multi-head self-attention and context-aware mechanisms to model temporal emotion progression across dialogue turns. Recent works like MemoCMT [43] further evolved this paradigm by applying cross-modal transformers with HuBERT–BERT embeddings, enabling long-range dependency modeling and fine-grained contextual synchronization. Through these refinements, attention-based fusion has established itself as a robust and interpretable design for modeling interdependent emotional cues.

Gating-based mechanisms focus on dynamically controlling the contribution of each modality to the final representation. Instead of fixed fusion weights, these methods learn adaptive gates—often implemented via sigmoid or softmax functions—that regulate the importance of the modality according to reliability, context, or signal quality. Representative studies such as COLD Fusion [24] and GMA [55] apply calibrated or gated attention modules that modulate fusion weights based on uncertainty estimation and modality-specific confidence. Similarly, DF-ERC and IF-MMIN [25] introduced dynamic gating to handle missing or degraded modalities, reconstructing latent representations through auxiliary imagination modules. These methods allow hybrid fusion models to adjust modality influence in real time, improving robustness under noise or partial observation. Gating mechanisms have thus become central to emotion recognition systems that operate under unpredictable environmental or conversational conditions.

Transformer-based and alignment-oriented hybrid fusion frameworks extend the attention paradigm by explicitly modeling the correspondence between modalities through positional encoding and multi-head projection spaces. These models treat each modality as a token sequence within a unified latent space, where self- and cross-attention jointly learn modality relationships across time. Architectures such as TDFNet [59], MGCMA [53], and CDaT [41] exemplify this approach, combining deep-scale transformers or contrastive alignment to capture hierarchical dependencies. More recent multimodal LLMs tuned to instruction, such as DialogueMLLM [30] and Emotion-LLaMA [29], further generalize the concept by embedding emotional cues from audio and visual modalities into language tokens, achieving unified reasoning through text-conditioned alignment. These frameworks represent the evolution of hybrid fusion toward scalable and semantically grounded architectures, merging foundation-model-level reasoning with traditional cross-modal interaction.

While hybrid fusion offers the most flexible and powerful integration paradigm, it often demands substantial computational resources and large-scale pretraining to achieve stable convergence. The use of multi-head attention, gating networks, and transformer stacks introduces additional parameters, leading to latency and interpretability trade-offs. To address these issues, recent research has explored lightweight hybrid designs such as ParallelNet [36] and SIA-Net [38], which maintain intermediate fusion capabilities through sparse attention and Shapley-value interpretability analysis. Collectively, hybrid fusion reflects the culmination of fusion strategy development, integrating the advantages of early and late paradigms while paving the way toward generalized, scalable, and context-aware multimodal emotion understanding.

2.4 Datasets and Evaluation Metrics

Datasets and evaluation metrics form the empirical backbone of MER research, shaping both model development and the way progress is assessed across studies. The choice of dataset determines the diversity of affective cues and interaction contexts available for learning, while evaluation metrics define how reliably these models capture emotional dynamics under varying conditions. This section first summarizes widely used multimodal emotion datasets and then reviews the primary metrics adopted in both categorical and continuous affect prediction.

Table 2: Summary of representative MER datasets from 2021–2025.

Dataset	Modalities	No. of Samples	Annotation
Audiovisual Datasets			
IEMOCAP [82]	V, A, T	12 h/10K utterances	Valence, Arousal, Categories
RAVDESS [83]	V, A	7356 clips	8 emotions
SAVEE [84]	V, A	480 clips	7 emotions
eNTERFACE [85]	V, A	1260 videos	6 emotions
CREMA-D [86]	V, A	7442 clips	6 emotions
MSP-IMPROV [87]	V, A, T	8438 segments	Valence, Arousal
BAUM-1 [88]	V, A	1200 videos	7 emotions
Aff-Wild2 [89]	V, A	2.8M frames	Valence, Arousal, Expr. Intensity

(Continued)

Table 2 (continued)

Dataset	Modalities	No. of Samples	Annotation
Text-Inclusive Multimodal Datasets			
CMU-MOSI [90]	V, A, T	2199 opinion segments	Sentiment/Valence
CMU-MOSEI [91]	V, A, T	23,453 segments	Sentiment/Emotions
MELD [92]	V, A, T	13,000 utterances	7 emotions
DailyDialog [93]	T, A	13,118 dialogues	7 emotions
EmoryNLP [94]	V, A, T	12,000 utterances	7 emotions
UR-FUNNY [95]	V, A, T	1864 videos	Humor + Emotion
MER2023 [96]	V, A, T	250 h/50K clips	8 emotions
MER2024 [97]	V, A, T	300 h (multilingual)	Valence, Arousal, Engagement
MER2025 [98]	V, A, T, S	500+ h/10K subjects	Valence, Arousal, Dominance
Physiological and Sensor-Based Datasets			
DEAP [99]	V, S	32 subjects × 40 trials	Valence, Arousal
MAHNOB-HCI [100]	V, S	30 subjects × 20 trials	Valence, Arousal, Dominance
BioVid EmoDB [101]	V, S	87 subjects × 900 trials	Intensity levels
AMIGOS [102]	V, S	40 subjects × 16 videos	Valence, Arousal, Liking
DREAMER [103]	S	23 subjects × 18 stimuli	Valence, Arousal, Dominance

Datasets

Datasets form the empirical foundation of MER, determining the richness of affective cues, the diversity of contexts, and the methodological direction of model development. The field has evolved from early, controlled audiovisual collections to large-scale, naturalistic datasets that integrate text, physiological signals, and conversational structures. This section outlines the major dataset categories used in MER research, highlighting their characteristics and the roles they play in shaping fusion strategies, representation learning, and evaluation protocols.

Early multimodal research relied primarily on audiovisual datasets, which provide aligned facial expressions and speech for studying foundational fusion architectures. Controlled and acted datasets such as RAVDESS [83], SAVEE [84], and eNTERFACE [85] offer clean, well-structured emotional expressions that are useful for benchmarking feature extraction and early/late fusion techniques. More complex audiovisual resources—including IEMOCAP [82], MSP-IMPROV [87], CREMA-D [86], and Aff-Wild2 [89]—introduce spontaneous interaction, continuous annotations, and environmental variability, enabling research on temporal modeling, robustness, and continuous affect prediction. Datasets such as BAUM-1 [88] further contribute to spontaneous interview-style emotional behavior, supporting studies on naturalistic visual–acoustic synchrony.

As multimodal learning expanded toward conversational and context-aware affect analysis, text-inclusive datasets became increasingly central. CMU-MOSI [90] and CMU-MOSEI [91] provide rich monologue-style annotations with corresponding audio and video streams, supporting research on sentiment grounding, multimodal alignment, and hybrid fusion. Dialogue-oriented datasets such as MELD [92], EmoryNLP [94], and UR-FUNNY [95] introduce multi-party interaction, turn-taking, and humor-related affect, expanding MER toward conversational settings. These datasets have been instrumental in advancing transformer-based architectures, cross-modal attention, and co-representation learning across text, audio, and vision.

In parallel, physiological and sensor-based datasets broadened the scope of MER by focusing on internal affective responses that are often inaccessible through external behavior alone. DEAP [99], MAHNOB-HCI [100], AMIGOS [102], BioVid EmoDB [101], and DREAMER [103] contain combinations of EEG, ECG, GSR, EMG, rPPG, and auxiliary video recordings. Their controlled designs and continuous valence–arousal annotations make them essential for studying affective computing from a biometric perspective, enabling research on domain adaptation, sensor fusion, and personalized emotion modeling. These datasets complement audiovisual corpora by capturing implicit affective states that remain stable even under occlusion or ambiguous facial behavior.

Building on these foundations, recent years have introduced large-scale and open-domain datasets designed to support robust MER in real-world environments. Resources such as MER2023 [96], MER2024 [97], and the MER2025 Challenge dataset [98] capture multilingual, cross-domain, and sensor-augmented emotional behavior across diverse demographics. Their large size and ecological variability enable the training of data-intensive models, such as transformer-based fusion networks, cross-modal contrastive learners, and self-supervised multimodal encoders, while providing benchmarks suitable for evaluating generalization beyond controlled laboratory settings.

A consolidated overview of representative MER datasets is presented in Table 2. Together, these datasets illustrate the progression from controlled audiovisual corpora to rich, multimodal, large-scale resources that reflect the complexity of real-world affect. This evolution has shaped the methodological direction of MER, enabling increasingly sophisticated fusion architectures and representation learning frameworks.

Evaluation Metrics

The evaluation of MER models depends on whether the task is defined as discrete emotion classification or continuous affect regression. Each formulation requires metrics that capture different properties of emotional expressiveness, class discriminability, and temporal reliability. This subsection formalizes the metrics most widely used in MER research and highlights representative studies that employ these measures.

Discrete emotion recognition is commonly used in datasets such as RAVDESS [83], IEMOCAP [82], CMU-MOSEI [91], and MELD [92]. Accuracy is the simplest and most frequently reported metric, defined as

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{N}, \tag{1}$$

where TP_i is the number of correctly predicted samples of class i, C is the total number of classes, and N is the total number of samples. Although used in early multimodal systems [65,71], accuracy is sensitive to class imbalance and therefore insufficient as a primary metric. Recent MER research instead emphasizes the macro-averaged F1-score, which assigns equal weight to each emotion category. For each emotion class i, precision and recall are defined as

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \qquad \text{Recall}_i = \frac{TP_i}{TP_i + FN_i}. \tag{2}$$

The class-wise F1-score is then computed as the harmonic mean of precision and recall as

$$\text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \tag{3}$$

Macro-F1 treats all emotion classes equally by averaging per-class F1-scores as

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i. \tag{4}$$

In contrast, Weighted-F1 adjusts each class contribution based on its sample proportion as

$$\text{Weighted-F1} = \sum_{i=1}^{C} w_i \cdot \text{F1}_i, \qquad w_i = \frac{N_i}{\sum_{j=1}^{C} N_j}. \tag{5}$$

where w_i is the number of samples in class i. Although informative, it may obscure poor performance on infrequent emotions, making macro-F1 more suitable for benchmarking.

Continuous affect estimation, used in Aff-Wild2 [89], DEAP [99], MAHNOB-HCI [100], and AMIGOS [102], requires metrics that account for temporal consistency and scale agreement between predicted and annotated emotional trajectories. The Concordance Correlation Coefficient (CCC) is the most widely adopted measure and is defined as

$$\text{CCC} = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{6}$$

where ρ_{xy} is the Pearson correlation coefficient, μ_x, μ_y are mean values, and σ_x, σ_y are standard deviations of predictions and ground truth. CCC penalizes both scale mismatch and mean shift, making it the official metric in Aff-Wild2 and ABAW challenges [89,104].

Mean squared error (MSE) provides a direct measure of the absolute difference between predicted and reference emotional trajectories. It emphasizes the magnitude of deviations at each timestep through a squared aggregation of residuals.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2, \tag{7}$$

where x_i denotes the predicted emotional value at timestep i, y_i denotes the corresponding ground-truth annotation, and N indicates the total number of temporal samples. MSE highlights pointwise discrepancy and is frequently adopted in continuous valence–arousal tasks [40,73].

Root mean squared error (RMSE) expresses prediction discrepancy in the same scale as the target signal, improving interpretability while retaining the sensitivity of squared errors.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}. \tag{8}$$

Variables follow the same definitions as in MSE, with RMSE providing a scale-aligned measure of overall prediction deviation [74].

Together, these metrics provide a consistent basis for evaluating MER systems across both categorical and continuous affect settings. Macro-F1 has become the predominant metric for discrete emotion classification because it reflects balanced class-wise performance, while CCC remains the standard in continuous emotion regression due to its combined assessment of correlation and agreement. These complementary metrics support reliable comparison across fusion strategies, embedding architectures, and cross-modal interaction designs in MER.

3 Future Research Directions

Despite significant progress in MER over recent years, current systems still face several foundational challenges that limit their robustness, scalability, and applicability to real-world environments. The core difficulty arises from the inherent heterogeneity of multimodal affective signals, which leads to persistent issues with temporal alignment, modality reliability, and cross-domain generalization. Because visual, acoustic, textual, and physiological cues evolve on different timescales and possess distinct noise characteristics, MER models often struggle to maintain consistent cross-modal interactions, particularly when signals are asynchronous or partially missing. Furthermore, varying levels of sensor noise, occlusion, background interference, and modality imbalance undermine the stability of fusion mechanisms, revealing a gap between controlled experimental settings and deployment scenarios. In addition, emotional expressions differ substantially across speakers, cultures, and situational contexts, creating a strong domain shift that current models are not yet equipped to handle. These challenges collectively highlight the need for new research directions that move beyond incremental architectural modifications toward more adaptive, generalizable, and reliability-aware multimodal learning frameworks. The following subsections outline emerging opportunities and potential pathways for advancing the next generation of MER systems.

3.1 Cross-Modal Alignment Challenges

One of the most persistent difficulties in MER lies in aligning heterogeneous signals that differ in temporal scale, structural properties, and semantic density. Visual cues evolve at the frame level, acoustic features exhibit continuous prosodic variation, and textual representations encapsulate discrete linguistic semantics. These discrepancies create a modality alignment gap, in which features extracted from different streams fail to correspond to the same emotional events in terms of time or semantic granularity. When such misalignment accumulates, fusion modules overfit to spurious correlations or suppress informative modality-specific cues, ultimately degrading affective inference.

Several recent studies highlight the consequences of this alignment gap across diverse multimodal settings. Audiovisual models often struggle to synchronize facial movements with rapidly varying speech prosody [68,70], while dialogue-level MER frameworks report instability when textual context shifts more slowly than speech or facial expressions [18,20]. Physiological–visual fusion pipelines further exaggerate this issue, as EEG or rPPG signals operate at substantially higher sampling rates than video frames [2,39]. Collectively, these findings indicate that alignment errors propagate through the network, increasing modality dominance, reducing complementarity, and weakening the reliability of downstream fusion.

To mitigate these effects, MER research has explored alignment-aware representation learning that explicitly constrains temporal or semantic correspondence before fusion. Common strategies include cross-modal attention mechanisms that dynamically match salient segments across modalities [19,60], alignment losses that encourage synchronized embedding trajectories through shared latent spaces [25,61], and hierarchical gating functions that suppress unsynchronized modality features during context modeling [20,55]. Frame-to-token matching and temporal interpolation layers have also been adopted to bridge differences in sampling density between visual and acoustic streams [26]. These approaches collectively indicate a gradual shift toward alignment-preserving fusion pipelines that prioritize correspondence before integration.

This alignment challenge extends beyond emotion recognition and arises broadly in multimodal learning scenarios involving heterogeneous and asynchronous data. In domains such as remote sensing and Earth observation, hybrid deep learning models are commonly used to fuse satellite imagery with auxiliary time-series signals, including meteorological or environmental measurements, where spatial observations and temporal dynamics are inherently misaligned [105,106]. Similar to MER, these frameworks integrate convolutional encoders for visual streams with recurrent or transformer-based modules for temporal signals [5,6], enabling intermediate-level alignment that preserves modality-specific structure while learning shared representations [1]. This parallel highlights that alignment-aware hybrid fusion constitutes a general architectural response to heterogeneous data integration rather than a domain-specific solution confined to affective computing.

Despite these advances, achieving consistent, fine-grained alignment remains an open challenge. High-frequency transitions in speech, rapid head movements, background noise, and variable speaking styles introduce non-stationarity, complicating temporal matching. Moreover, current alignment methods often rely on fixed attention windows or pairwise matching, limiting their ability to model long-range emotional dependencies. Future progress requires scalable alignment mechanisms that integrate temporal dynamics, semantic abstraction, and uncertainty modeling into a unified framework, enabling multimodal systems to capture emotional correspondence with higher precision and stability.

3.2 Modality Reliability and Robust Fusion

In real-world MER, modalities are often missing or degraded due to environmental noise, occlusion, motion blur, sensor artifacts, or transcription errors. Audio streams may suffer from low signal-to-noise ratios or overlapping speakers, visual frames are often affected by rapid head movements or lighting changes, and textual transcripts derived from automatic speech recognition can contain alignment errors. Physiological signals, such as EEG and rPPG, also fluctuate significantly across sessions and subjects. These inconsistencies lead to modality imbalance, where one modality becomes unreliable or unavailable during inference, resulting in a substantial drop in recognition performance. As MER systems increasingly move toward unconstrained, in-the-wild deployment, developing architectures that maintain stable performance under partial or noisy multimodal input has become an essential research direction.

To address this challenge, several frameworks incorporate mechanisms for learning robust representations that remain informative even when certain modalities are incomplete or corrupted. Modality-invariant feature learning, exemplified by IF-MMIN, improves robustness by aligning distributional characteristics across modalities and generating plausible latent representations when specific inputs are missing [25]. Graph-based models such as MGAFR refine cross-modal correlations using multiplex graph aggregation and contrastive refinement, offering stable performance in unsupervised and incomplete-modality scenarios [51]. Diffusion-driven restoration approaches, including RMER-DT, reconstruct missing modalities through denoising-based generation before applying hierarchical transformer fusion [44]. In physiological–behavioral settings, CMSLNet integrates adaptive consistency metrics and joint attention to achieve cross-subject robustness with heterogeneous EEG and eye-tracking signals [40]. Reliability-aware fusion mechanisms have also emerged; GMA introduces gated cross-modal enhancement to filter redundant signals [55], and CDaT dynamically adjusts the contribution of each modality by transferring information from more reliable streams to less reliable ones [41]. These approaches collectively demonstrate the increasing importance of modeling modality confidence, cross-modal redundancy, and latent restoration.

Despite these advances, achieving high robustness under severe modality degradation remains challenging. Generating latent substitutes for missing modalities introduces the risk of over-smoothing or hallucinated emotional cues, especially when the remaining modalities provide limited contextual evidence. Reliability estimation often fluctuates across speakers, domains, and recording conditions, leading to unstable weighting during fusion. Physiological signals are susceptible to noise and personalization effects, making it challenging to generalize subjects even when adaptive normalization is applied. Furthermore, most existing methods assume that at least one high-quality modality is available during inference; handling situations where all modalities are partially corrupted remains difficult. These limitations highlight the need for more principled formulations of modality uncertainty and cross-modal redundancy.

Future research may address these issues by developing generative reconstruction frameworks that integrate diffusion models, masked autoencoding, or modality-specific priors to restore incomplete inputs more faithfully. Another promising direction involves self-supervised reliability estimation, where modality confidence is learned without explicit labels by predicting consistency across temporal neighborhoods or cross-modal agreement. Cross-modal redundancy learning, in which the model identifies semantically shared information across modalities, can further reduce dependence on any single stream. Large multimodal language models may also support self-correction through internally generated descriptions or synthetic complementary signals. More comprehensive uncertainty modeling, spanning epistemic, aleatoric, and cross-modal uncertainty, can guide adaptive fusion policies that generalize across diverse real-world settings. These directions collectively represent an important step toward building MER systems capable of maintaining stable performance in incomplete, noisy, and operationally unconstrained environments.

3.3 Generalization across Speakers and Environments

MER models frequently experience substantial performance degradation when deployed across speakers, cultures, and recording environments that differ from their training conditions. Variations in age, gender, speaking style, facial morphology, linguistic background, and sensor quality create distribution shifts that challenge conventional supervised learning pipelines. Even models trained on large-scale in-the-wild corpora often overfit to dataset-specific affective cues, limiting their generalization to new domains such as clinical interviews, automotive cabins, call-center conversations, or multi-speaker social interactions. These issues become more pronounced when modalities exhibit unequal robustness across domains—for instance, when visual cues degrade under low-light conditions or when acoustic features vary due to accent, microphone quality, and environmental noise.

Recent studies have sought to address these limitations through domain generalization and cross-subject adaptation strategies. Adversarial domain alignment has been employed to learn domain-invariant affective representations, for example, in cross-corpus speech emotion recognition and cross-subject EEG–video fusion models for emotion recognition [107,108]. Meta-learning has also been explored for rapid adaptation to unseen domains in text and speech emotion classification, improving generalization across datasets and label distributions [109]. Cross-cultural continuous emotion recognition frameworks further highlight the need for culture-aware normalization and calibration when transferring models across populations [110]. In parallel, GAN-based data augmentation has been used to synthesize additional emotional samples and increase robustness to recording conditions [111]. Alongside these modeling techniques, zero-shot and cross-domain evaluation protocols have emerged in large-scale multimodal LLM-based systems, where models such as Emotion-LLaMA are tested on unseen corpora without task-specific fine-tuning [29]. Although these approaches collectively reduce sensitivity to domain-specific biases, substantial gaps persist between in-domain and zero-shot performance, especially under significant demographic or environmental shifts.

Promising research directions include developing pretraining pipelines compatible with emerging foundation models and drawing on large-scale, multilingual, and multicultural affective corpora to capture broad emotional variability. Another direction is the integration of parameter-efficient adaptation modules such as low-rank adaptation [112] and prompt-based modulation, which enable general-purpose multimodal encoders to specialize in emotion-related reasoning without extensive retraining. Expanding the coverage of speaker and environment diversity through weak supervision or pseudo-labeled affective data also offers a practical path toward more robust generalization. In addition, developing context-aware representations that capture environmental factors, lighting conditions, background noise, and social interaction patterns can reduce a model's reliance on identity-specific cues and encourage the extraction of stable, transferable emotional signals. These advances point toward a future in which MER systems achieve reliable performance across diverse demographic, linguistic, and situational contexts suited for deployment in real-world environments.

4 Conclusion

MER has evolved into a central research field within human-centered artificial intelligence, driven by the need to understand complex affective cues that span visual, acoustic, textual, and physiological modalities. This survey consolidated recent advances across feature representation, fusion methodologies, and dataset development, highlighting how multimodal architectures have transitioned from early concatenation-based frameworks to hybrid fusion strategies that leverage cross-modal attention, contrastive alignment, and transformer-based interaction modeling. The review further demonstrated that the rapid growth of pre-trained encoders and large-scale multimodal datasets has substantially expanded the design space of affective modeling, enabling more expressive and context-aware emotional representations.

Despite these advances, significant challenges remain, including asynchronous modality alignment, incomplete or noisy signals, domain and demographic generalization, and the integration of large multimodal foundation models. The future research directions discussed in this survey emphasize scalable pretraining pipelines, parameter-efficient adaptation, context-aware representation learning, and robust multimodal synchronization frameworks. Collectively, these developments suggest a clear trajectory toward unified, adaptive, and deployment-ready emotion recognition models that operate reliably across diverse environments.

Acknowledgement: The authors acknowledge that this research was conducted in connection with IITP-funded projects supported by the Korea government (MSIT) (Grant Nos. RS-2021-II211341 and 2021-0-00766).

Funding Statement: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation grant funded by the Korea government (MSIT) (No. RS-2021-II211341, AI Graduate School Support Program, Chung-Ang University), and in part by the Institute of Information and Communications Technology Planning and Evaluation grant funded by the Korea government (MSIT) (Development of Integrated Development Framework that Supports Automatic Neural Network Generation and Deployment Optimized for Runtime Environment, Grant No. 2021-0-00766).

Author Contributions: The authors confirm contribution to the paper as follows: A-Seong Moon contributed to the conceptualization, validation, formal analysis, project administration, original draft preparation, review and editing, and secured key resources for the study. Haesung Kim contributed to the methodology design, data curation, investigation, and assisted in the preparation of the original manuscript draft. Ye-Chan Park contributed to the validation, visualization, and participated in the review and editing of the manuscript. Jaesung Lee supervised the overall research process, contributed to the conceptualization and project administration. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Summary of Recent MER Studies

This appendix provides a comprehensive summary table of representative multimodal emotion recognition studies published between 2021 and 2025.

Table A1: Summary of representative MER studies published from 2021 to 2025.

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[43]	2025	A, T	Hybrid Fusion	A: HuBERT T: BERT	IEMOCAP, ESD, MELD	Introduces a cross-modal transformer fusion that strengthens audio–text alignment and improves conversational MER.
[47]	2025	V, A, T	Early Fusion	V: CMU-SDK A: CMU-SDK T: BERT	CMU-MOSEI	Applies XAI-based feature selection to identify influential multimodal cues, improving interpretability in MER.
[44]	2025	V, A, T	Hybrid Fusion	V: DenseNet A: OpenSmile T: RoBERTa	IEMOCAP, MELD	Combines diffusion-based restoration and hierarchical Transformers to improve robustness in conversational MER.
[45]	2025	V, A, T	Early Fusion	V: 3D-CNN A: OpenSmile T: RoBERTa	IEMOCAP, MELD	Examines MER through graph spectrum theory to enhance high-frequency affective signal preservation.
[46]	2025	V, A, T	Late Fusion	V: ResNet-50 A: Semi-CNN T: DeBERTa	BAUM-1, SAVEE	Integrates text, speech, face, and motion signals using modular late fusion for comprehensive MER.
[48]	2025	V, S	Early Fusion	V: MTCNN S: pre-processing	DEAP	Combines EEG and facial expressions to capture both physiological and observable cues for emotion discrimination.
[50]	2025	A, T	Late Fusion	A: CARE T: RoBERTa	IEMOCAP, MELD, CMU-MOSI	Uses LLM-supervised pretraining to improve semantic alignment and contextual emotion understanding.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[51]	2025	V, A, T	Hybrid Fusion	V: MA-Net A: Wav2vec T: DeBERTa	CMU-MOSI, CMU-MOSEI	Introduces multiplex graph aggregation to address incomplete modalities and refine multimodal feature relations.
[79]	2025	V, S	Hybrid Fusion	V: 1D-CNN S: 1D-CNN	RECOLA, ULM-TSST	Proposes dual-view disentanglement and hierarchical reconstruction to improve robustness under modality noise.
[113]	2025	V, A, T	Late Fusion	V: OpenFace A: COVERAP T: T5	Aff-Wild2	Develops a dynamic-scene-aware MER framework robust to real-world variations and complex environments.
[76]	2025	V, A, T	Early Fusion	V: ViT A: GNN T: CapsNet	MELD, CMU-MOSEI	Employs capsule–graph Transformer architecture to enhance relational modeling of multimodal affective cues.
[53]	2025	A, T	Hybrid Fusion	A: Wav2vec2 T: BERT	IEMOCAP	Introduces multi-granularity cross-modal alignment for better synchronization of emotional cues.
[52]	2025	V, A, T, S	Hybrid Fusion	V: CLIP A: Wav2vec2 T: CLIP S: ViT	MELD	Uses spiking Transformers to improve robustness against adversarial attacks in multimodal emotion recognition.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[55]	2025	V, A, T	Hybrid Fusion	V: DenseNet A: OpenSmile T: RoBERTa	IEMOCAP, MELD	Applies cross-modal gating to enhance interaction strength and refine multimodal feature learning.
[54]	2025	V, A, T, S	Late Fusion	V, A, T, S: Transformer	ASCERTAIN, KEMDy20	Introduces a gated mixture-of-experts architecture with cross-attention for multimodal and sensor-based MER.
[30]	2025	V, A, T	Hybrid Fusion	V,A,T: Video-LLaMA	MELD	Introduces an instruction-tuned multimodal LLM that enhances conversational emotion reasoning and fusion robustness.
[49]	2025	V, A	Late Fusion	V: MTCNN, EfficientNet A: Wav2vec2	CREMA-D, eNTERFACE	Develops an efficient audiovisual pipeline that leverages deep speech and facial encoders for robust MER.
[64]	2025	V, A	Hybrid Fusion	V: VGG A: X-vector	RAVDESS, SAVEE, CREMA-D	Combines audiovisual temporal modeling with attention-based fusion to capture dynamic emotional transitions.
[29]	2024	V, A, T	Early Fusion	V: ViT A: HuBERT T: LLaMA2	MERR, MER2023, MER2024, DFEW, EMER	Proposes an instruction-tuned multimodal LLaMA that unifies perception and reasoning for large-scale MER.
[77]	2024	V, A, T	Early Fusion	V: CNN A: 1D Conv T: Bi-LSTM	MELD, IEMOCAP	Addresses class imbalance using deep imbalanced learning and cross-modal integration for conversational MER.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[65]	2024	V, A	Hybrid Fusion	V: CNN A: CNN	BAUM-1, RAVDESS	Evaluates multiple fusion schemes and shows that combined CNN-based audiovisual cues improve affective recognition.
[34]	2024	A, T	Early Fusion	A: CNN T: BERT	MELD, CMU-MOSEI	Incorporates attention-enhanced BERT–CNN fusion to improve linguistic–acoustic alignment in MER tasks.
[31]	2024	V, A, T	Hybrid Fusion	V: 3D-CNN A: Opensmile T: RoBERTa	IEMOCAP, MELD	Introduces masked graph learning with recurrent alignment for stable cross-modal fusion in dialogue MER.
[66]	2024	V, A	Early Fusion	V: PyFEAT A: Opensmile	RAVDESS, BAUM-1	Leverages LLM-based fusion of handcrafted visual and acoustic cues to enhance multimodal representation learning.
[56]	2024	V, A, T	Late Fusion	V: CNN A: 1D CNN T: fastText	CMU-MOSEI	Designs a lightweight late-fusion pipeline optimized for real-time MER with competitive performance.
[35]	2024	V, A, T	Hybrid Fusion	V, A, T: MLP	CMU-MOSI, CMU-MOSEI, CH-SIMS	Presents a token-disentangling transformer that isolates modality-specific and shared cues for improved MER.
[32]	2024	V, A, T	Hybrid Fusion	V: 3D-CNN A: Opensmile T: RoBERTa	IEMOCAP, MELD	Applies adversarial alignment and information bottleneck fusion to improve robustness in conversational MER.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[114]	2024	V, A, T	Early, Late Fusion	V: Py-Feat A: Wav2vec2 T: Whisper	C-EXPR-DB, MELD	Compares textualized and feature-based MER, showing advantages of converting modalities into rich text embeddings.
[2]	2024	V, S	Early Fusion	V: DSCNN S: Bi-LSTM	BioVid EmoDB	Introduces a healthcare-oriented MER framework integrating physiological and video cues with model-level fusion.
[81]	2024	V, A, T	Late Fusion	V: DenseNet A: Opensmile, COVAREP T: BERT	IEMOCAP, MSP-IMPROV, CMU-MOSI	Uses contrastive learning to acquire modality-invariant features for missing-modality MER scenarios.
[75]	2024	V, A, T	Hybrid Fusion	V, A: Transformer T: BLIP	MET-Meme, MemeCap	Introduces metaphor-aware alignment using context disentangling to improve robustness in multimodal emotion recognition.
[40]	2024	V, S	Hybrid Fusion	V, S: MLP	SEED-IV, SEED-V	Presents a multisource learning framework that enhances cross-subject stability through comprehensive modal integration.
[78]	2024	A, T	Hybrid Fusion	A: Wav2vec2 T: BERT	IEMOCAP	Develops fine-grained disentangled representation learning to better separate emotional factors across modalities.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[36]	2024	V, A	Hybrid Fusion	V: VGG A: VGG	IIT-R SIER	Proposes an interpretable fusion pipeline combining visual and speech cues for transparent multimodal emotion analysis.
[37]	2024	A, T	Hybrid Fusion	A: HuBERT T: KoELECTRA	KER (AI-HUB)	Integrates Korean speech and text encoders with a multimodal transformer for culturally aligned MER modeling.
[58]	2024	V, A, T	Hybrid Fusion	V: DenseNet A: Opensmile T: RoBERTa	IEMOCAP, MELD	Introduces persona-infused graph learning to capture emotion shifts and conversational structure across modalities.
[67]	2024	V, A	Late Fusion	V, A: CNN	CREMA-D, RAVDESS	Uses uncertainty-based weighting to optimize a lightweight audiovisual model for efficient emotion recognition.
[39]	2024	V, S	Hybrid Fusion	V: ResNet S: MTTS-CAN	MAHNOB-HCI	Combines facial expression cues with rPPG signals to build an end-to-end physiological–visual emotion recognizer.
[42]	2024	V, A, T	Early Fusion	V: DenseNet, SVM A: Opensmile T: RoBERTa	IEMOCAP, MELD	Proposes calibration for conversational MER through early fusion and modality-specific reliability adjustments.
[74]	2024	V, S	Hybrid Fusion	V: ResNet S: DGC	DEAP, SEED-IV	Applies dense GCN with joint cross-attention to enhance emotional state inference from video and physiological signals.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[115]	2024	V, A, T	Hybrid Fusion	V: 3D-CNN A: Opensmile T: TextCNN	IEMOCAP, MELD	Incorporates speaker-aware cognitive modeling with cross-modal attention to better capture dialogue-based emotions.
[41]	2024	V, A, T	Hybrid Fusion	V: LSTM A: LSTM T: BERT	CMU-MOSEI, IEMOCAP	Introduces dynamic transfer learning across modalities to enhance generalization in multimodal emotion tasks.
[38]	2024	V, A, T	Early Fusion	V: FabNet A: Wav-RoBERTa T: RoBERTa	MELD, CMU-MOSI, CMU-MOSEI	Presents sparse interactive attention enabling effective integration of fine-grained cues from all modalities.
[33]	2024	V, A, T	Hybrid Fusion	V: DenseNet A: Opensmile T: RoBERTa	IEMOCAP, MELD, DailyDialog	Models relational subgraph interactions to capture emotion-dependent patterns across diverse modalities.
[116]	2024	V, A, T	Early Fusion	V: DenseNet, 3D-CNN A: Opensmile T: RoBERTa	IEMOCAP	Introduces reinforcement learning–based optimization to refine early-fusion representations for MER tasks.
[73]	2023	V, S	Hybrid Fusion	V: DeepVANet S: CNN	DEAP, MAHNOB-HCI	Combines EEG and facial expression cues with hybrid fusion to enhance emotional state estimation in multimodal settings.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[117]	2023	V, A, S	Late Fusion	V: GhostNet A: LFCNN S: tLSTM	CK+, EMO-DB, MAHNOB-HCI	Integrates facial, speech, and EEG modalities via late fusion to improve multimodal affect recognition robustness.
[18]	2023	V, A, T	Hybrid Fusion	V: DenseNet A: Opensmile T: RoBERTa	MELD, IEMOCAP	Revisits disentanglement of modality and context to refine conversational MER with enhanced fusion strategies.
[21]	2023	V, A, T	Early Fusion	V, A: CNN, Transformer T: ALBERT	IEMOCAP, CMU-MOSEI	Proposes transformer-based early fusion with emotion-level representation learning for improved multi-label MER.
[19]	2023	A, T	Hybrid Fusion	A: Bi-LSTM T: BERT	MELD, IEMOCAP	Uses hybrid attention networks to effectively integrate audio and text cues for conversational emotion analysis.
[20]	2023	V, A, T	Hybrid Fusion	V: DenseNet A: Opensmile T: RoBERTa	MELD, IEMOCAP	Introduces a transformer model with self-distillation to strengthen multimodal learning in dialogue contexts.
[27]	2023	V, A	Hybrid Fusion	V: 3D ResNet A: ResNet	CREMA-D, RAVDESS	Employs cross-modal audio–video fusion with attention and metric learning to enhance emotion classification.
[22]	2023	V, A, T	Hybrid Fusion	V: VGG A: CNN T: DialogXL	IEMOCAP	Provides a unified evaluation framework for comparing multimodal fusion techniques across emotion tasks.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[25]	2023	V, A, T	Hybrid Fusion	V: DenseNet A: Opensmile T: BERT	IEMOCAP	Learns modality-invariant representations to achieve robust emotion recognition under missing modality conditions.
[23]	2023	V, A, T	Hybrid Fusion	V: expMAE A: HuBERT T: MacBERT	MER-SEMI	Introduces semi-supervised learning with expression MAE and multi-modal pseudo-labeling to enhance MER accuracy.
[26]	2023	V, A	Early Fusion	V: ResNet A: Transformer	RAVDESS, eNTERFACE	Combines facial and speech features through early fusion to improve emotion prediction in audiovisual settings.
[59]	2023	A, T	Late Fusion	A: BuBERT, ECAPA T: BERT	IEMOCAP	Proposes TDFNet for deep-scale fusion of audio and text signals using transformer modules for emotion inference.
[118]	2023	V, A, T	Hybrid Fusion	V: ResNet A: HuBERT T: MacBERT	MER2023-SEMI	Uses class-balanced pseudo-labeling to enhance multimodal learning under limited supervision conditions.
[24]	2023	V, A	Hybrid Fusion	V: EmoFAN A: VGGish	AVEC, CMU-MOSEI, IEMOCAP	Introduces calibrated ordinal latent distribution fusion to model uncertainty in multimodal emotion recognition.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[60]	2023	A, T	Hybrid Fusion	A: Wav2vec2 T: BERT	IEMOCAP	Improves fusion of Wav2vec2 and BERT by incorporating auxiliary tasks that strengthen cross-modal alignment.
[61]	2023	A, T	Hybrid Fusion	A: Wav2vec2 T: BERT	IEMOCAP	Introduces a Bayesian co-attention mechanism that leverages knowledge-aware priors to strengthen audio–text fusion for MER.
[119]	2023	A, T	Hybrid Fusion	A: Wav2vec2 T: FlauBERT	CEMO	Analyzes attention mechanisms for emotion detection in emergency-call speech by modeling cross-modal audio–text dependencies.
[28]	2023	A, T	Hybrid Fusion	A: CNN, BiGRU T: BiGRU	IEMOCAP	Uses deep temporal modeling and cross-modal transformers to capture synchronized emotional cues from audio and text streams.
[68]	2022	V, A	Late Fusion	V: DSN, DTN A: CNN	SAVEE, RAVDESS, RML	Proposes a spatio-temporal CNN framework integrating facial and vocal features through late fusion for emotion recognition.
[12]	2022	V, A, T	Hybrid Fusion	V: OpenFace, Facet A: COVAREP T: BERT	CMU-MOSI, CMU-MOSEI, UR-FUNNY	Disentangles modality-specific and contextual features to improve robustness and fusion effectiveness in conversational MER.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[69]	2022	V, A	Late Fusion	V, A: CNN	RAVDESS, SAVEE	Applies model-level late fusion using CNN-based visual and acoustic encoders to improve audiovisual emotion classification.
[62]	2022	V, A, T	Hybrid Fusion	V: 3D CNN A: CNN-GRU T: Bi-LSTM	IEMOCAP	Integrates speech, video, and MoCAP cues using hybrid fusion to model fine-grained temporal dynamics in emotional behavior.
[13]	2022	V, A, T	Hybrid Fusion	V: 3D CNN A: Opensmile T: CNN	IEMOCAP, AVEC, MELD	Models hierarchical uncertainty across modalities to improve robustness of multimodal emotion understanding in conversations.
[120]	2022	V, A	Early Fusion	V, A: -	SAVEE, eNTER-FACE	Uses K-means-guided kernel CCA to enhance early-fusion alignment in human–robot interaction emotion recognition tasks.
[14]	2022	V, A, T	Hybrid Fusion	V: DenseNet A: Wav2vec2 T: BERT	IEMOCAP, MSP-IMPROV	Introduces a prompt-based pretraining strategy that enhances multimodal alignment for improved downstream MER performance.
[16]	2022	V, S	Late Fusion	V: ResNet S: CNN	Bio Vid Emo DB, CIFE	Combines facial features and affective biomarkers to support emotion-state prediction in smart-industry health analytics.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[80]	2022	V, A, T	Hybrid Fusion	V: ResNet A: CNN T: Word2Vec	IEMOCAP, MSP-IMPROV	Employs a weight-sharing autoencoder with multi-head attention to enhance multimodal feature refinement and fusion.
[15]	2022	V, S	Hybrid Fusion	V: CNN S: 1D CNN	In-house database	Combines facial cues with wireless sensing signals to provide a multimodal perspective for emotion detection applications.
[57]	2022	V, A	Hybrid Fusion	V: R(2+1)D, ST-GCN A: TCN T: Transformer	RAVDESS, CMU-MOSEI	Uses pairwise contrastive loss to enforce modality consistency and strengthen cross-modal representation learning.
[17]	2022	A, T	Hybrid Fusion	V: FAN A: PANNs, LEAF T: Transformer	IEMOCAP, CMU-MOSEI	Explores cross-modal translation to leverage additional datasets, improving generalization in audio–text MER settings.
[8]	2021	V, A, T	Hybrid Fusion	V: Facet A: COVAREP T: Glove	CMU-MOSI, CMU-MOSEI, IEMO-CAP	Introduces progressive reinforcement to handle unaligned streams and enhance multimodal consistency in emotional cues.
[72]	2021	V, S	Hybrid Fusion	V: CNN S: SVM	FER2013, SEED-IV	Combines facial features with EEG-based affective signals to improve robustness in multimodal emotion classifications.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[70]	2021	V, A	Late Fusion	V: OpenFace A: Wav2vec2	RAVDESS	Applies aural transformers and facial action units with late fusion to improve audiovisual emotion detection accuracy.
[9]	2021	V, A, T	Late Fusion	V: FaceNet, GRU A: WaveRNN T: GPT	MELD	Uses transformer-based cross-modal fusion to achieve robust conversational MER under complex emotional variability.
[63]	2021	A, T	Hybrid Fusion	A: log-MFB T: BERT	IEMOCAP	Models temporal–semantic consistency to enhance alignment between acoustic changes and text-based emotional cues.
[71]	2021	V, A	Hybrid Fusion	V: STN A: CNN	RAVDESS	Uses transfer learning pipelines for audiovisual analysis, improving generalization in limited-size emotional datasets.
[10]	2021	V, A, T	Hybrid Fusion	V: VGG A: VGGish, SoundNet T: BERT	CMU-MOSI, CMU-MOSEI, IEMOCAP	Unifies heterogeneous features using BERT-driven fusion, enhancing representation quality across three modalities.
[121]	2021	V, A, T	Hybrid Fusion	V: ResNet T: CNN, Bi-GRU	WikiArt Emotions	Introduces sequential co-attention to analyze emotional signals in artworks by combining visual and semantic cues.

(Continued)

Table A1 (continued)

Ref.	Year	Modality	Approach	Feature Extraction	Datasets	Contribution
[122]	2021	V, A	Hybrid Fusion	V: VGG A: CNN	eNTERFACE	Uses capsule graph convolution to refine audiovisual features and strengthen multimodal emotion representation.
[11]	2021	V, A, T	Hybrid Fusion	V, A, T: Bi-GRU	IEMOCAP, CMU-MOSI	Applies deep CCA-based fusion to extract correlated multimodal features and improve stability in emotional predictions.
[123]	2021	V, A	Hybrid Fusion	V: SIFT, CNN A: PyAudio	Friends dataset	Integrates handcrafted visual features and speech descriptors using deep belief networks for emotion recognition.
[124]	2021	V, A, T	Early Fusion	V: FEA A: AcousEmo T: CrystalFeel	OMG Emotion	Analyzes contextual and environmental factors in video-based MER, emphasizing early fusion of multimodal cues.

References

1. Wu Y, Mi Q, Gao T. A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions. Biomimetics. 2025;10(7):418. doi:10.3390/biomimetics10070418.
2. Islam MM, Nooruddin S, Karray F, Muhammad G. Enhanced multimodal emotion recognition in healthcare analytics: a deep learning based model-level fusion approach. Biomed Signal Process Control. 2024;94(1):106241. doi:10.1016/j.bspc.2024.106241.
3. Zhou H, Liu Z. Realization of self-adaptive higher teaching management based upon expression and speech multimodal emotion recognition. Front Psychol. 2022;13:857924. doi:10.3389/fpsyg.2022.857924.
4. Chen Z, Feng X, Zhang S. Emotion detection and face recognition of drivers in autonomous vehicles in IoT platform. Image Vis Comput. 2022;128:104569. doi:10.1016/j.imavis.2022.104569.
5. Ramaswamy MPA, Palaniswamy S. Multimodal emotion recognition: a comprehensive review, trends, and challenges. Wiley Interdiscip Rev Data Min Know Disc. 2024;14(6):e1563. doi:10.1002/widm.1563.
6. Kalateh S, Estrada-Jimenez LA, Nikghadam-Hojjati S, Barata J. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. IEEE Access. 2024;12:103976–4019. doi:10.1109/access.2024.3430850.
7. Chang KC, Chen SQ. A survey of multimodal emotion recognition: fusion techniques, datasets, challenges and future directions. Int J Biomet. 2025;17(5):485–510. doi:10.1504/ijbm.2025.148281.

8. Lv F, Chen X, Huang Y, Duan L, Lin G. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Virtual. p. 2554–62. doi:10.1109/cvpr46437.2021.00258.
9. Xie B, Sidulova M, Park CH. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. Sensors. 2021;21(14):4913. doi:10.3390/s21144913.
10. Lee S, Han DK, Ko H. Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. IEEE Access. 2021;9:94557–72. doi:10.1109/access.2021.3092735.
11. Zhang K, Li Y, Wang J, Wang Z, Li X. Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis. IEEE Signal Process Letters. 2021;28:1898–902. doi:10.1109/lsp.2021.3112314.
12. Yang D, Huang S, Kuang H, Du Y, Zhang L. Disentangled representation learning for multimodal emotion recognition. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisbon, Portugal. p. 10–4.
13. Chen F, Shao J, Zhu A, Ouyang D, Liu X, Shen HT. Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. IEEE Trans Cybern. 2022;54(1):187–98. doi:10.1109/tcyb.2022.3185119.
14. Zhao J, Li R, Jin Q, Wang X, Li H. MEmoBERT: pre-training model with prompt-based learning for multimodal emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ, USA: IEEE; 2022. p. 4703–7.
15. Dang X, Chen Z, Hao Z, Ga M, Han X, Zhang X, et al. Wireless sensing technology combined with facial expression to realize multimodal emotion recognition. Sensors. 2022;23(1):338. doi:10.3390/s23010338.
16. Kumar A, Sharma K, Sharma A. MEmoR: a multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries. Image Vis Comput. 2022;123:104483. doi:10.1016/j.imavis.2022.104483.
17. Yoon YC. Can we exploit all datasets? Multimodal emotion recognition using cross-modal translation. IEEE Access. 2022;10:64516–24. doi:10.1109/access.2022.3183587.
18. Li B, Fei H, Liao L, Zhao Y, Teng C, Chua TS, et al. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. p. 5923–34. doi:10.1145/3581783.3612053.
19. Zhang S, Yang Y, Chen C, Liu R, Tao X, Guo W, et al. Multimodal emotion recognition based on audio and text by using hybrid attention networks. Biomed Signal Process Control. 2023;85:105052. doi:10.1016/j.bspc.2023.105052.
20. Ma H, Wang J, Lin H, Zhang B, Zhang Y, Xu B. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Trans Multim. 2023;26:776–88. doi:10.1109/tmm.2023.3271019.
21. Le HD, Lee GS, Kim SH, Kim S, Yang HJ. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. IEEE Access. 2023;11:14742–51. doi:10.1109/access.2023.3244390.
22. Peña D, Aguilera A, Dongo I, Heredia J, Cardinale Y. A framework to evaluate fusion methods for multimodal emotion recognition. IEEE Access. 2023;11:10218–37. doi:10.1109/access.2023.3240420.
23. Cheng Z, Lin Y, Chen Z, Li X, Mao S, Zhang F, et al. Semi-supervised multimodal emotion recognition with expression mae. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. p. 9436–40. doi:10.1145/3581783.3612840.
24. Tellamekala MK, Amiriparian S, Schuller BW, André E, Giesbrecht T, Valstar M. COLD fusion: calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. IEEE Trans Pattern Anal Mach Intell. 2023;46(2):805–22. doi:10.1109/tpami.2023.3325770.
25. Zuo H, Liu R, Zhao J, Gao G, Li H. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 Jun 4–10; Rhodes, Greece. p. 1–5.
26. Tang G, Xie Y, Li K, Liang R, Zhao L. Multimodal emotion recognition from facial expression and speech based on feature fusion. Multim Tools Appl. 2023;82(11):16359–73. doi:10.1007/s11042-022-14185-0.
27. Mocanu B, Tapu R, Zaharia T. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image Vis Comput. 2023;133:104676. doi:10.1016/j.imavis.2023.104676.

28. Maji B, Swain M, Guha R, Routray A. Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 Jun 4–10; Rhodes, Greece. p. 1–5.
29. Cheng Z, Cheng ZQ, He JY, Wang K, Lin Y, Lian Z, et al. Emotion-LLaMA: multimodal emotion recognition and reasoning with instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems. London, UK; 2024. Vol. 37, p. 110805–53.
30. Sun Y, Zhou T. DialogueMLLM: transforming multimodal emotion recognition in conversation through instruction-tuned MLLM. IEEE Access. 2025;13:121048–60. doi:10.1109/access.2025.3591447.
31. Meng T, Zhang F, Shou Y, Shao H, Ai W, Li K. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. IEEE/ACM Trans Audio Speech Lang Process. 2024;32:4298–312. doi:10.1109/taslp.2024.3434495.
32. Shou Y, Meng T, Ai W, Zhang F, Yin N, Li K. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. Inf Fusion. 2024;112:102590. doi:10.1016/j.inffus.2024.102590.
33. Wang Y, Zhang W, Liu K, Wu W, Hu F, Yu H, et al. Dynamic emotion-dependent network with relational subgraph interaction for multimodal emotion recognition. IEEE Trans Affect Comput. 2025;16(2):712–25. doi:10.1109/taffc.2024.3461148.
34. Makhmudov F, Kultimuratov A, Cho YI. Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures. Appl Sci. 2024;14(10):4199. doi:10.20944/preprints202404.1574.v1.
35. Yin G, Liu Y, Liu T, Zhang H, Fang F, Tang C, et al. Token-disentangling mutual transformer for multimodal emotion recognition. Eng Appl Artif Intell. 2024;133:108348. doi:10.1016/j.engappai.2024.108348.
36. Kumar P, Malik S, Raman B. Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. Multim Tools Appl. 2024;83(10):28373–94. doi:10.1007/s11042-023-16443-1.
37. Yi MH, Kwak KC, Shin JH. KoHMT: a multimodal emotion recognition model integrating KoELECTRA, HuBERT with multimodal transformer. Electronics. 2024;13(23):4674. doi:10.3390/electronics13234674.
38. Li S, Zhang T, Chen CP. Sia-net: sparse interactive attention network for multimodal emotion recognition. IEEE Trans Computat Soc Syst. 2024;11(5):6782–94. doi:10.1109/tcss.2024.3409715.
39. Li J, Peng J. End-to-end multimodal emotion recognition based on facial expressions and remote photoplethysmography signals. IEEE J Biomed Health Inform. 2024;28(10):6054–63. doi:10.1109/jbhi.2024.3430310.
40. Chen C, Li Z, Kou KI, Du J, Li C, Wang H, et al. Comprehensive multisource learning network for cross-subject multimodal emotion recognition. IEEE Trans Emerg Topics Computat Intell. 2024;9(1):365–80. doi:10.1109/tetci.2024.3406422.
41. Hong S, Kang H, Cho H. Cross-modal dynamic transfer learning for multimodal emotion recognition. IEEE Access. 2024;12:14324–33. doi:10.1109/access.2024.3356185.
42. Tu G, Xiong F, Liang B, Wang H, Zeng X, Xu R. Multimodal emotion recognition calibration in conversations. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 9621–30.
43. Khan M, Tran PN, Pham NT, El Saddik A, Othmani A. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. Sci Rep. 2025;15(1):5473. doi:10.1038/s41598-025-89202-x.
44. Zhu X, Wang Y, Cambria E, Rida I, López JS, Cui L, et al. RMER-DT: robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. Inf Fusion. 2025;123(C):103268. doi:10.1016/j.inffus.2025.103268.
45. Ai W, Zhang F, Shou Y, Meng T, Chen H, Li K. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2025 Feb 25–Mar 4; Philadelphia, PA, USA. p. 11418–26.
46. Boitel E, Mohasseb A, Haig E. MIST: multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis. Expert Syst Appl. 2025;270:126236. doi:10.1016/j.eswa.2024.126236.
47. Khalane A, Makwana R, Shaikh T, Ullah A. Evaluating significant features in context-aware multimodal emotion recognition with XAI methods. Expert Syst. 2025;42(1):e13403. doi:10.22541/au.167407909.97031004/v1.

48. Güler SE, Akbulut FP. Multimodal emotion recognition: emotion classification through the integration of EEG and facial expressions. IEEE Access. 2025;13(1):24587–603. doi:10.1109/access.2025.3538642.
49. Fan S, Jing J, Wang C. Audio-visual learning for multimodal emotion recognition. Symmetry. 2025;17(3):418. doi:10.3390/sym17030418.
50. Dutta S, Ganapathy S. LLM supervised pre-training for multimodal emotion recognition in conversations. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2025 Apr 6–11; Hyderabad, India. p. 6–11.
51. Deng Y, Bian J, Wu S, Lai J, Xie X. Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition. Inf Fusion. 2025;114(1):102711. doi:10.1016/j.inffus.2024.102711.
52. Chen G, Qian Z, Zhang D, Qiu S, Zhou R. Enhancing robustness against adversarial attacks in multimodal emotion recognition with spiking transformers. IEEE Access. 2025;13:34584–97. doi:10.1109/access.2025.3544086.
53. Wang X, Zhao S, Sun H, Wang H, Zhou J, Qin Y. Enhancing multimodal emotion recognition through multi-granularity cross-modal alignment. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2025 Apr 6–11; Hyderabad, India. p. 6–11.
54. Mengara Mengara AG, Moon Y-K. CAG-MoE: multimodal emotion recognition with cross-attention gated mixture of experts. Mathematics. 2025;13(12):1907. doi:10.3390/math13121907.
55. Zhao S, Ren J, Zhou X. Cross-modal gated feature enhancement for multimodal emotion recognition in conversations. Sci Rep. 2025;15(1):30004. doi:10.1038/s41598-025-11989-6.
56. Dixit C, Satapathy SM. Deep CNN with late fusion for real time multimodal emotion recognition. Expert Syst Appl. 2024;240(1):122579. doi:10.1016/j.eswa.2023.122579.
57. Franceschini R, Fini E, Beyan C, Conti A, Arrigoni F, Ricci E. Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. In: Proceedings of the 26th International Conference on Pattern Recognition; 2022 Aug 21–25; Montreal, QC, Canada. p. 2589–96.
58. Tu G, Xiong F, Liang B, Xu R. A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024 Jul 14–18; Washington, DC, USA. p. 14–8.
59. Zhao Z, Wang Y, Shen G, Xu Y, Zhang J. TDFNet: transformer-based deep-scale fusion network for multimodal emotion recognition. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:3771–82. doi:10.1109/taslp.2023.3316458.
60. Sun D, He Y, Han J. Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 Jun 4–10; Rhodes, Greece. p. 1–5.
61. Zhao Z, Wang Y, Wang Y. Knowledge-aware bayesian co-attention for multimodal emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 Jun 4–10; Rhodes, Greece. p. 1–5.
62. Jia N, Zheng C, Sun W. A multimodal emotion recognition model integrating speech, video and MoCAP. Multim Tools Appl. 2022;81(22):32265–86. doi:10.1007/s11042-022-13091-9.
63. Chen B, Cao Q, Hou M, Zhang Z, Lu G, Zhang D. Multimodal emotion recognition with temporal and semantic consistency. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3592–603. doi:10.1109/taslp.2021.3129331.
64. Salas-Cáceres J, Lorenzo-Navarro J, Freire-Obregón D, Castrillón-Santana M. Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. Multim Tools Appl. 2025;84(23):27327–43. doi:10.1007/s11042-024-20227-6.
65. Bilotti U, Bisogni C, De Marsico M, Tramonte S. Multimodal emotion recognition via convolutional neural networks: comparison of different strategies on two multimodal datasets. Eng Appl Artif Intell. 2024;130(1):107708. doi:10.1016/j.engappai.2023.107708.
66. Chandraumakantham O, Gowtham N, Zakariah M, Almazyad A. Multimodal emotion recognition using feature fusion: an LLM-based approach. IEEE Access. 2024;12:108052–71. doi:10.1109/access.2024.3425953.
67. Radoi A, Cioroiu G. Uncertainty-based Learning of a lightweight model for multimodal emotion recognition. IEEE Access. 2024;12:120362–74. doi:10.1109/access.2024.3450674.

68. Sharafi M, Yazdchi M, Rasti R, Nasimi F. A novel spatio-temporal convolutional neural framework for multimodal emotion recognition. Biomed Signal Process Control. 2022;78:103970. doi:10.1016/j.bspc.2022.103970.
69. Middya AI, Nag B, Roy S. Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities. Knowl Based Syst. 2022;244(3):108580. doi:10.1016/j.knosys.2022.108580.
70. Luna-Jiménez C, Kleinlein R, Griol D, Callejas Z, Montero JM, Fernández-Martínez F. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. Appl Sci. 2021;12(1):327. doi:10.3390/app12010327.
71. Luna-Jiménez C, Griol D, Callejas Z, Kleinlein R, Montero JM, Fernández-Martínez F. Multimodal emotion recognition on RAVDESS dataset using transfer learning. Sensors. 2021;21(22):7665. doi:10.3390/s21227665.
72. Tan Y, Sun Z, Duan F, Solé-Casals J, Caiafa CF. A multimodal emotion recognition method based on facial expressions and electroencephalography. Biomed Signal Process Control. 2021;70:103029. doi:10.1016/j.bspc.2021.103029.
73. Wang S, Qu J, Zhang Y, Zhang Y. Multimodal emotion recognition from EEG signals and facial expressions. IEEE Access. 2023;11:33061–8. doi:10.1109/access.2023.3263670.
74. Cheng C, Liu W, Feng L, Jia Z. Dense graph convolutional with joint cross-attention network for multimodal emotion recognition. IEEE Trans Computat Soc Syst. 2024;11(5):6672–83. doi:10.1109/tcss.2024.3412074.
75. Zhang L, Jin L, Xu G, Li X, Xu C, Wei K, et al. CAMEL: capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024 Feb 20–27; Vancouver, BC, Canada. p. 9341–9.
76. Filali H, Boulealam C, El Fazazy K, Mahraz AM, Tairi H, Riffi J. Meaningful multimodal emotion recognition based on capsule graph transformer architecture. Information. 2025;16(1):40. doi:10.3390/info16010040.
77. Meng T, Shou Y, Ai W, Yin N, Li K. Deep imbalanced learning for multimodal emotion recognition in conversations. IEEE Trans Artif Intell. 2024;5(12):6472–87. doi:10.1109/tai.2024.3445325.
78. Sun H, Zhao S, Wang X, Zeng W, Chen Y, Qin Y. Fine-grained disentangled representation learning for multimodal emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2024 Apr 14–19; Seoul, Republic of Korea. p. 11051–5.
79. Li C, Xie L, Wang X, Pan H, Wang Z. A twin disentanglement Transformer Network with Hierarchical-Level Feature Reconstruction for robust multimodal emotion recognition. Expert Syst Appl. 2025;264(5):125822. doi:10.1016/j.eswa.2024.125822.
80. Zheng J, Zhang S, Wang Z, Wang X, Zeng Z. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. IEEE Trans Multim. 2022;25:2213–25. doi:10.1109/tmm.2022.3144885.
81. Liu R, Zuo H, Lian Z, Schuller BW, Li H. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. IEEE Trans Affect Comput. 2024;15(4):1856–73. doi:10.1109/taffc.2024.3378570.
82. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Evaluat. 2008;42(4):335–59. doi:10.1007/s10579-008-9076-6.
83. Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One. 2018;13(5):e0196391. doi:10.32920/25412950.v1.
84. Jackson P, Haq S. Surrey audio-visual expressed emotion (savee) database. Guildford, UK: University of Surrey; 2014.
85. Martin O, Kotsia I, Macq B, Pitas I. The eNTERFACE'05 audio-visual emotion database. In: Proceedings of the 22nd International Conference on Data Engineering Workshops; 2006 Apr 3–7; Atlanta, GA, USA.
86. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREAM-D: crowd-sourced emotional multimodal actors dataset. IEEE Trans Affect Comput. 2014;5(4):377–90. doi:10.1109/taffc.2014.2336244.
87. Busso C, Parthasarathy S, Burmania A, AbdelWahab M, Sadoughi N, Provost EM. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. IEEE Trans Affect Comput. 2016;8(1):67–80. doi:10.1109/taffc.2016.2515617.

88. Zhalehpour S, Onder O, Akhtar Z, Erdem CE. BAUM-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans Affect Comput. 2016;8(3):300–13. doi:10.1109/taffc.2016.2553038.
89. Kollias D, Zafeiriou S. Aff-Wild2: extending the aff-wild database for affect recognition. arXiv:1811.07770. 2018.
90. Zadeh A, Zellers R, Pincus E, Morency LP. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259. 2016.
91. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, VIC, Australia. p. 2236–46.
92. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 527–36.
93. Li Y, Su H, Shen X, Li W, Cao Z, Niu S. DailyDialog: a manually labelled multi-turn dialogue dataset. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017 Nov 27–Dec 1; Taipei, Taiwan. p. 986–95.
94. Zahiri SM, Choi JD. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In: Proceedings of the AAAI Workshops; 2018 Feb 2–3; New Orleans, LA, USA. p. 44–52.
95. Hasan MK, Rahman W, Zadeh AB, Zhong J, Tanveer MI, Morency LP, et al. UR-FUNNY: a multimodal language dataset for understanding humor. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019 Nov 3–7; Hong Kong, China. p. 2046–56.
96. Lian Z, Sun H, Sun L, Chen K, Xu M, Wang K, et al. MER 2023: multi-label learning, modality robustness, and semi-supervised learning. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. p. 9610–4.
97. Lian Z, Sun H, Sun L, Wen Z, Zhang S, Chen S, et al. MER 2024: semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In: Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 41–8.
98. Lian Z, Liu R, Xu K, Liu B, Liu X, Zhang Y, et al. MER 2025: when affective computing meets large language models. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; San Francisco, CA, USA. p. 13837–42.
99. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, et al. Deap: a database for emotion analysis; using physiological signals. IEEE Trans Affect Comput. 2011;3(1):18–31. doi:10.1109/t-affc.2011.15.
100. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput. 2011;3(1):42–55. doi:10.1109/t-affc.2011.25.
101. Zhang L, Walter S, Ma X, Werner P, Al-Hamadi A, Traue HC, et al. “BioVid Emo DB”: a multimodal database for emotion analyses validated by subjective ratings. In: Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence; 2016 Dec 6–9; Athens, Greece. p. 1–6.
102. Miranda-Correa JA, Abadi MK, Sebe N, Patras I. AMIGOS: a dataset for affect, personality and mood research on individuals and groups. IEEE Trans Affect Comput. 2018;12(2):479–93. doi:10.1109/taffc.2018.2884461.
103. Katsigiannis S, Ramzan N. DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inform. 2017;22(1):98–107. doi:10.1109/jbhi.2017.2688239.
104. Kollias D. ABAW: valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 2328–36.
105. Lin F, Crawford S, Guillot K, Zhang Y, Chen Y, Yuan X, et al. MMST-VIT: climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2–6; Paris, France. p. 5774–84.

106. Mohamed SA, Maksoud OOA, Fathy A, Mohamed AS, Hosny K, Keshk HM, et al. A hybrid deep learning and rule-based model for smart weather forecasting and crop recommendation using satellite imagery. Sci Rep. 2025;15(1):36102. doi:10.1038/s41598-025-21506-4.

107. Feng K, Chaspari T. A review of generalizable transfer learning in automatic emotion recognition. Front Comput Sci. 2020;2:9. doi:10.3389/fcomp.2020.00009.

108. Guo Y, Tang C, Wu H, Chen B. GNN-based multi-source domain prototype representation for cross-subject EEG emotion recognition. Neurocomputing. 2024;609(3):128445. doi:10.1016/j.neucom.2024.128445.

109. Tang Y, Wu Y, He Y, Liu J, Zhang W. Dual-task contrastive meta-learning for few-shot cross-domain emotion recognition. Comput Mater Contin. 2025;82(2):2331–52. doi:10.32604/cmc.2024.059115.

110. Gong X, Chen CP, Zhang T. Cross-cultural emotion recognition with EEG and eye movement signals based on multiple stacked broad learning system. IEEE Trans Computat Soc Syst. 2023;11(2):2014–25. doi:10.1109/tcss.2023.3298324.

111. Shilandari A, Marvi H, Khosravi H, Wang W. Speech emotion recognition using data augmentation method by cycle-generative adversarial networks. Signal Image Video Process. 2022;16(7):1955–62. doi:10.1007/s11760-022-02156-9.

112. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2022.

113. Liu L, Luo Q, Zhang W, Zhang M, Zhai B. Multimodal emotion recognition method in complex dynamic scenes. J Inform Intell. 2025;3(3):257–74. doi:10.1016/j.jiixd.2025.02.004.

114. Richet N, Belharbi S, Aslam H, Schadt ME, González-González M, Cortal G, et al. Textualized and feature-based models for compound multimodal emotion recognition in the wild. In: Proceedings of the European Conference on Computer Vision; 2024 Sep 29–Oct 4; Milan, Italy. p. 60–78.

115. Guo L, Song Y, Ding S. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. Knowl Based Syst. 2024;296(3):111969. doi:10.1016/j.knosys.2024.111969.

116. Zhang C, Zhang Y, Cheng B. RL-EMO: a reinforcement learning framework for multimodal emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2024 Apr 14–19; Seoul, Republic of Korea. p. 10246–50.

117. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal emotion recognition based on facial expressions, speech, and EEG. IEEE Open J Eng Med Biol. 2023;5:396–403. doi:10.1109/ojemb.2023.3240280.

118. Chen H, Guo C, Li Y, Zhang P, Jiang D. Semi-supervised multimodal emotion recognition with class-balanced pseudo-labeling. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 29–Nov 3; Ottawa, ON, Canada. p. 9556–60.

119. Deschamps-Berger T, Lamel L, Devillers L. Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2023 Jun 4–10; Rhodes, Greece. p. 1–5.

120. Chen L, Wang K, Li M, Wu M, Pedrycz W, Hirota K. K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction. IEEE Trans Indust Electr. 2022;70(1):1016–24. doi:10.1109/tie.2022.3150097.

121. Tashu TM, Hajiyeva S, Horvath T. Multimodal emotion recognition from art using sequential co-attention. J Imaging. 2021;7(8):157. doi:10.3390/jimaging7080157.

122. Liu J, Chen S, Wang L, Liu Z, Fu Y, Guo L, et al. Multimodal emotion recognition with capsule graph convolutional based representation fusion. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2021 Jun 6–11; Toronto, ON, Canada. p. 6339–43.

123. Liu D, Chen L, Wang Z, Diao G. Speech expression multimodal emotion recognition based on deep belief network. J Grid Comput. 2021;19(2):22. doi:10.1007/s10723-021-09564-0.

124. Bhattacharya P, Gupta RK, Yang Y. Exploring the contextual factors affecting multimodal emotion recognition in videos. IEEE Trans Affect Comput. 2021;14(2):1547–57. doi:10.1109/taffc.2021.3071503.