A Comprehensive Review in Unimodal and Multimodal Emotion Recognition

JIACHEN LUO, Centre for Digital Music, Queen Mary University of London, United Kingdom and Technical University of Munich, Germany
QU YANG*, Tencent, China
JIAJUN HE*, Information Technology Center, Nagoya University, Japan
YINING HUA*, Harvard University, USA
ZHENG LIAN*, State Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, China
YUANCHAO LI*, The University of Edinburgh, United Kingdom
SIYANG SONG*, University of Exeter, United Kingdom
WEN WU*, Shanghai Artificial Intelligence Laboratory, China
DINGDONG WANG*, The Chinese University of Hong Kong, China
SHUAI SHEN*, Nanyang Technological University, Singapore
JINGYAO WU*, Media Lab, Massachusetts Institute of Technology, USA
GUIMIN HU*, Guangdong University of Technology, China
HE HU*, Shenzhen University, China
YONG LI*, Southeast University, China
ZIXING ZHANG*, Hunan University, China
JIADONG WANG*, Technical University of Munich, Germany
SIFAN ZHOU*, Carnegie Mellon University, USA
ZUOJIN TANG*, Zhejiang University, China
CANRAN XIAO*, Sun Yat-sen University, China
SHENG XU*, The Chinese University of Hong Kong, Shenzhen, China
ZHENJUN ZHAO*, The Chinese University of Hong Kong, China and University of Zaragoza, Spain
XIANGYANG XUE*, Fudan University, China
SICHENG ZHAO*, Tsinghua University, China
YONG DAI*, Beijing Innovation Center of Humanoid Robotics, China
TOMOKI TODA*, Information Technology Center, Nagoya University, Japan
LICAI SUN*, University of Oulu, Finland
KAILAI YANG, The University of Manchester, United Kingdom
LIYUN ZHANG, The University of Tokyo, Japan
CONG CAI, University of Chinese Academy of Sciences, China
JIAMIN DU, University of Oxford, United Kingdom
ZIYANG MA, Shanghai Jiao Tong University, China
MINGJIE CHEN, University of Sheffield, United Kingdom
CHENGXUAN QIAN, Texas A&M University, USA
ZHENLONG YUAN, University of California, Santa Cruz, USA
XIE CHEN, Shanghai Jiao Tong University, China
HUY PHAN†, Meta Reality Lab, USA
LIN WANG†, Centre for Digital Music, Queen Mary University of London, United Kingdom
BJOERN SCHULLER†, Imperial College London, United Kingdom and Technical University of Munich, Germany
JOSHUA REISS†, Centre for Digital Music, Queen Mary University of London, United Kingdom

Authors' addresses: Jiachen Luo, jiachen.luo@qmul.ac.uk, Centre for Digital Music, Queen Mary University of London, London, United Kingdom and Technical University of Munich, Munich, Germany; Qu Yang*, yangqu@whu.edu.cn, Tencent, Shenzhen, China; Jiajun He*, jiajun.he@g.sp.m.is.nagoya-u.ac.jp, Information Technology Center, Nagoya University, Nagoya, Japan; Yining Hua*, yininghua@g.harvard.edu, Harvard University, Cambridge, MA, USA; Zheng Lian*, lianzheng@tongji.edu.cn, State Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai, China; Yuanchao Li*, yuanchao.li@ed.ac.uk, The University of Edinburgh, Edinburgh, United Kingdom; Siyang Song*, s.song@exeter.ac.uk, University of Exeter, Exeter, United Kingdom; Wen Wu*, wuwen@pjlab.org.cn, Shanghai Artificial Intelligence Laboratory, Shanghai, China; Dingdong Wang*, dingdongwang1@gmail.com, The Chinese University of Hong Kong, Hong Kong, China; Shuai Shen*, shensthu19@gmail.com, Nanyang Technological University, Singapore, Singapore; Jingyao Wu*, jingyaow@mit.edu, Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA; Guimin Hu*, rice.hu.x@gmail.com, Guangdong University of Technology, Guangzhou, China; He Hu*, huhe@gml.ac.cn, Shenzhen University, Shenzhen, China; Yong Li*, mysee1989@gmail.com, Southeast University, Nanjing, China; Zixing Zhang*, zixingzhang@hnu.edu.cn, Hunan University, Changsha, China; Jiadong Wang*, jiadong.wang@tum.de, Technical University of Munich, Munich, Germany; Sifan Zhou*, sifanjay@gmail.com, Carnegie Mellon University, Pittsburgh, PA, USA; Zuojin Tang*, zuojintang@zju.edu.cn, Zhejiang University, Hangzhou, China; Canran Xiao*, xiaocr3@mail.sysu.edu.cn, Sun Yat-sen University, Guangzhou, China; Sheng Xu*,

Emotion recognition is a fundamental component of human-centered intelligent systems, supporting applications in healthcare, education, marketing, and human–computer interaction. Despite rapid progress driven by deep learning across facial, speech, textual, and multi-modal settings, the literature remains difficult to compare due to inconsistent emotion models, heterogeneous datasets, and varying evaluation protocols. This survey addresses this gap by providing a unified synthesis of deep learning-based uni-modal and multi-modal emotion recognition within a coherent analytical framework covering emotion modeling, dataset curation, representation learning, fusion strategies, and evaluation. Rather than listing methods, we organize existing work around key structural choices and trade-offs that affect generalization. For uni-modal approaches, we analyze how facial, speech, and textual methods increasingly rely on self-supervised pretraining to mitigate annotation scarcity, while retaining modality-specific limitations. For multi-modal systems, we examine alignment, modality dominance, complementarity, robustness, and the emerging role of large language models in affective reasoning. We further highlight persistent challenges, including label ambiguity, cross-dataset generalization, fairness, and the gap between benchmark performance and real-world deployment. This survey provides a unified perspective and a roadmap for future research. Resources are available at `https://github.com/jackchen69/Awesome-Emotion-Models`.

CCS Concepts: • **affective computing, deep learning, multi-modal emotion recognition, multimodality, emotion models**;

1 INTRODUCTION

Automatic emotion recognition is an increasingly prominent research area aimed at enabling machines to understand human affective states. Early studies primarily focused on single-modality analysis, such as facial, speech, or textual emotion recognition [1–8]. Facial expressions, such as smiling or lifting the eyebrows, often indicate emotions like happiness or surprise. Prosodic characteristics in speech, including changes in pitch or loudness, can similarly reveal states such as sadness or anger. Textual content also conveys affective information through word choice, sentence structure, and contextual meaning, offering valuable insight into an individual's emotional state. However, the accuracy of single-modal emotion recognition is often limited because each modality captures only a partial view of human affect and is easily disrupted by noise. These limitations motivated researchers to incorporate multiple modalities in order to achieve more robust and comprehensive emotional assessments, ultimately paving the way for the transition from early single-channel studies to more integrated multi-modal approaches. (see Fig. 1).

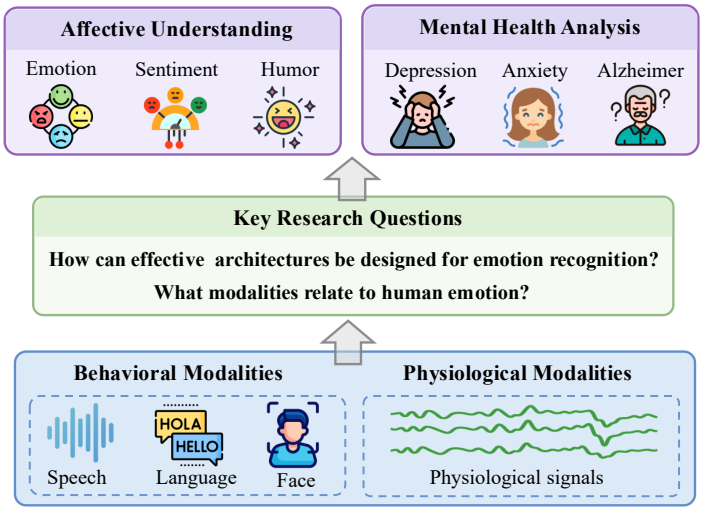


Fig. 1. Humans express emotions through behavioral and physiological modalities. Emotion recognition has potential applications in healthcare, multimedia, human-computer interaction, and robotics, among other areas.

Emotion is inherently multi-modal: both internal physiological responses and external behavioral cues contribute to how affect is expressed and perceived. Multi-modal emotion recognition leverages this richness by integrating heterogeneous sources of information, such as speech, text, facial expressions, and physiological signals, in order to achieve higher accuracy and robustness than uni-modal approaches [3, 4, 9–11]. By combining complementary cues across modalities, multi-modal emotion recognition systems can better handle noise, occlusion, and linguistic ambiguity, which often degrade performance when relying on a single modality [12, 13]. As illustrated in Fig. 2, multi-modal fusion enables a more fine-grained interpretation of affective states, improving both the estimation of emotion intensity and the consistency of categorical labels across speaker turns. Consequently, MER has become a central focus in affective computing, offering a more comprehensive and reliable framework for emotional inference.

shengxu1@link.cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China; Zhenjun Zhao*, ericzzj89@gmail.com, The Chinese University of Hong Kong, Hong Kong, China and University of Zaragoza, Zaragoza, Spain; Xiangyang Xue*, xyxue@fudan.edu.cn, Fudan University, Shanghai, China; Sicheng Zhao*, schzhao@tsinghua.edu.cn, Tsinghua University, Beijing, China; Yong Dai*, daiyongya@outlook.com, Beijing Innovation Center of Humanoid Robotics, Beijing, China; Tomoki Toda*, tomoki@icts.nagoya-u.ac.jp, Information Technology Center, Nagoya University, Nagoya, Japan; Licai Sun*, licai.sun@oulu.fi, University of Oulu, Oulu, Finland; Kailai Yang, klyang990203@gmail.com, The University of Manchester, Manchester, United Kingdom; Liyun Zhang, liyun.zhang@lab.ime.cmc.osaka-u.ac.jp, The University of Tokyo, Tokyo, Japan; Cong Cai, caicong@bit.edu.cn, University of Chinese Academy of Sciences, Beijing, China; Jiamin Du, jiamin.du@psych.ox.ac.uk, University of Oxford, Oxford, United Kingdom; Ziyang Ma, zym.22@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Mingjie Chen, mingjie.chen@sheffield.ac.uk, University of Sheffield, Sheffield, United Kingdom; Chengxuan Qian, open.qiancx@gmail.com, Texas A&M University, College Station, TX, USA; Zhenlong Yuan, zyuan54@ucsc.edu, University of California, Santa Cruz, Santa Cruz, CA, USA; Xie Chen, chenxie95@sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Huy Phan†, huyphan@meta.com, Meta Reality Lab, Menlo Park, CA, USA; Lin Wang†, lin.wang@qmul.ac.uk, Centre for Digital Music, Queen Mary University of London, London, United Kingdom; Bjoern Schuller†, schuller@tum.de, Imperial College London, London, United Kingdom and Technical University of Munich, Munich, Germany; Joshua Reiss†, joshua.reiss@qmul.ac.uk, Centre for Digital Music, Queen Mary University of London, London, United Kingdom.

With the rapid advancement of deep learning methodologies, emotion recognition based on both unimodal and multimodal data has become a major research focus in this field. A key challenge in this area lies in the effective design of neural network architectures and corresponding loss functions. Deep learning models are particularly well-suited for emotion recognition tasks because of their ability to automatically learn discriminative feature representations and to integrate information from multiple modalities. In recent years, considerable attention has been devoted to multi-modal fusion strategies, especially those incorporating attention mechanisms, among which scaled dot-product attention has gained widespread adoption. Motivated by these developments, this survey provides a comprehensive review of deep learning-based approaches for unimodal and multimodal emotion recognition and further outlines several promising directions for future research.

Despite substantial progress in emotion recognition research [3, 5, 14], surveys on unimodal emotion recognition have made notable contributions by examining individual modalities in depth—covering facial expression recognition, speech emotion recognition, and textual sentiment analysis, with detailed analyses of preprocessing pipelines, modality-specific feature representations, and model architectures [15–19]. However, these reviews share a fundamental limitation: they are largely confined to within-modality benchmarking and rarely align emotion label spaces, data collection conditions, or evaluation protocols across modalities. As a result, cross-modality comparisons remain difficult, and the field lacks a unified theoretical framework for synthesizing findings—since emotion annotations are inherently subjective and influenced by cultural, linguistic, and contextual factors, resulting in label ambiguity and poor cross-dataset or cross-language generalization.

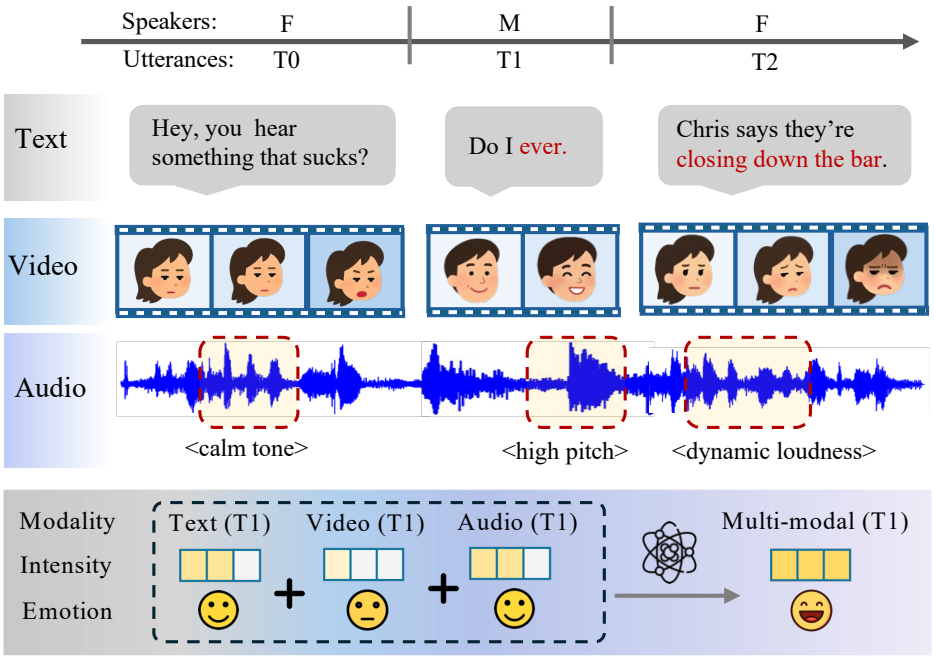


Fig. 2. Uni-modal vs. multi-modal emotion recognition in conversations. Video, Audio, and text capture among other facial, prosodic, and semantic cues. Speaker turns (F: female, M: male) are shown on the timeline. Multi-modal fusion yields more accurate emotional states across turns, including intensity and categorical labels.

Table 1. Comparison of our review with state-of-the-art surveys on uni-modal and multi-modal emotion recognition in conversation (2020–2026). A = Audio, T = Text, V = Visual, P = Physiological.

Pub. [Ref]	Year	Modality	Uni-modal	Multimodal	Evaluation	Pipeline	Dataset
Speech Commun [13]	2020	A	✓	✗	✓	✗	✓
IEEE TAFFC [14]	2020	A	✓	✗	✗	✗	✗
Information Fusion [5]	2020	A, T, V	✗	✓	✗	✗	✓
Electronics [20]	2021	A, T, V	✓	✓	✗	✓	✓
IEEE Signal Process. Mag. [21]	2021	A, T, V	✗	✓	✗	✗	✓
Information Science [22]	2022	A, T, V	✓	✗	✗	✗	✓
Neurocomputing [23]	2022	A, T, V	✗	✓	✗	✗	✓
IJACSA [24]	2022	A, T, V	✗	✓	✗	✗	✗
CMPB [25]	2022	A, T, V	✗	✓	✗	✗	✗
Information Fusion [26]	2022	A, T, V	✓	✓	✗	✗	✓
IEEE TIM [2]	2023	V	✓	✗	✗	✗	✓
Proc. IEEE [27]	2023	V	✓	✗	✓	✗	✓
IEEE TAFFC [28]	2023	T	✓	✗	✗	✗	✓
Speech Commun [29]	2023	A	✓	✗	✗	✗	✓
Neurocomputing [30]	2023	A	✓	✗	✗	✗	✓
Speech Commun [31]	2023	A	✓	✗	✗	✗	✓
ISWA [32]	2023	A	✓	✗	✗	✗	✓

(Continued on next page)

(Continued from previous page)

Pub. [Ref]	Year	Modality	Uni-modal	Multimodal	Evaluation	Pipeline	Dataset
IEEE Access [16]	2023	A	✓	✗	✗	✗	✓
Speech Commun [33]	2023	A, T	✗	✓	✗	✗	✓
ISWA [34]	2023	A, T, V	✗	✓	✗	✗	✓
Sensors [35]	2023	A, T, V	✗	✓	✗	✗	✓
Information Fusion [17]	2023	A, T, V	✓	✓	✗	✗	✓
Entropy [18]	2023	A, T, V	✓	✓	✓	✓	✓
Neurocomputing [36]	2023	A, T, V, P	✓	✓	✗	✗	✓
Information Fusion [1]	2024	V	✓	✗	✗	✗	✓
Information [37]	2024	V	✓	✗	✗	✗	✓
Speech Commun [38]	2024	A	✓	✗	✗	✗	✓
Information Fusion [9]	2024	A, T, V, P	✗	✓	✗	✗	✓
IEEE Access [10]	2024	A, T, V	✗	✓	✗	✗	✓
Information Fusion [11]	2024	A, T, V	✗	✓	✗	✗	✓
Expert Syst. Appl. [12]	2024	A, T, V	✓	✓	✗	✗	✓
Artificial Intelligence Review [37]	2025	A, T, V, P	✗	✓	✗	✗	✗
Expert Systems [39]	2025	A, T, V	✓	✓	✗	✗	✓
ACM TOMM [40]	2025	A, T, V, P	✗	✓	✗	✗	✓
IEEE Access [41]	2025	A, T, V	✓	✓	✗	✗	✓
ACM Trans. Interact. Intell. Syst. [42]	2026	P	✓	✗	✗	✗	✓
Information Fusion [43]	2026	S	✓	✗	✓	✓	✓
Stat. Optim. Inf. Comput. [44]	2026	F	✓	✗	✗	✗	✓
Ours	2026	A, T, V, P	✓	✓	✓	✓	✓

More recent surveys have shifted their attention toward multimodal emotion recognition, exploring how heterogeneous modalities can be jointly leveraged for richer affective understanding [3, 9, 12, 28, 45]. Despite this broader scope, these works share several critical shortcomings: most provide only coarse-grained taxonomies of fusion strategies without systematically examining how fusion design choices relate to modality availability, missing-modality robustness, or training objectives; evaluation practices remain fragmented across datasets, label schemes, and metrics, making reproducible comparison difficult; and the rapid emergence of large language models and multi-modal foundation models as new paradigms for affective reasoning is not adequately reflected. Consequently, the field still lacks a unified analytical framework spanning the full pipeline—from emotion modeling and data preparation to learning strategies, fusion architectures, and standardized evaluation.

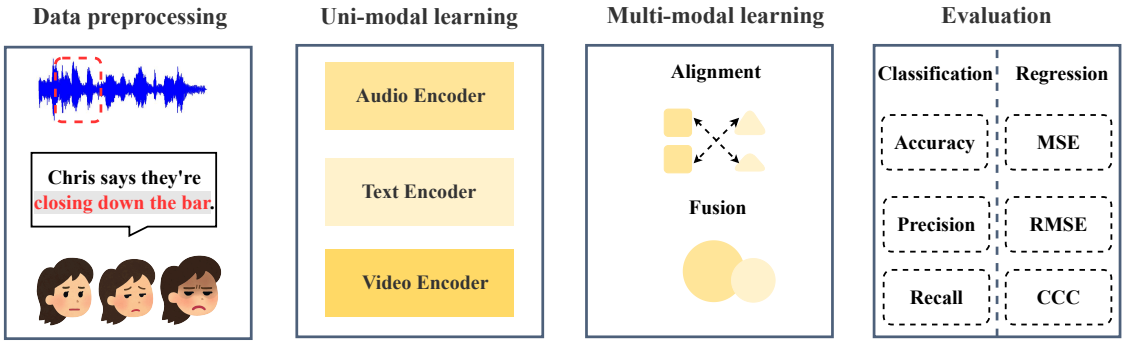


Fig. 3. We present a standardized implementation of a comprehensive multi-modal emotion recognition framework for conversational data, encompassing all critical components: data preprocessing, uni-modal learning, multi-modal learning, optimization, and metrics.

Motivated by the above analysis, this survey addresses three core gaps: (1) uni-modal surveys remain siloed within individual modalities and fail to provide a cross-modality analytical framework that aligns emotion label spaces, data conditions, and evaluation protocols; (2) multi-modal surveys offer only coarse-grained overviews of fusion strategies and overlook critical factors such as missing-modality robustness, training objectives, and standardized benchmarking; and (3) the growing influence of large-scale pre-trained and multi-modal foundation models on both unimodal and multimodal emotion recognition has not been systematically examined within a unified perspective.

To bridge these gaps, this survey jointly examines uni-modal and multi-modal paradigms under a coherent analytical framework that spans the full learning pipeline—from emotion modeling and dataset curation to feature learning, fusion architecture design, and evaluation. Beyond modeling and fusion strategies, our review explicitly addresses the often-overlooked roles of data preprocessing, emotion label design, and cross-dataset evaluation protocols, thereby enabling fair comparison and reproducible analysis across studies. We further highlight emerging trends such as benchmark unification, robust cross-domain and cross-linguistic evaluation, and the transformative impact of multi-modal foundation models on the field. Specifically, the main contributions of this survey are summarized as follows (see Fig. 4):

- **Deep Analytical Framework:** We introduce a structured taxonomy (Fig. 3) that provides a consistent lens for categorizing and analyzing emotion recognition studies across core dimensions, including data preprocessing, input representations, uni-modal learning, multi-modal fusion, and evaluation strategies. This framework harmonizes terminology and analytical criteria across the survey.
- **Systematic Synthesis of Uni-modal and Multi-modal Literature:** Guided by the proposed taxonomy, we systematically organize and compare existing works on deep uni-modal and multi-modal emotion recognition, as summarized in Table 1. Each component of the framework is

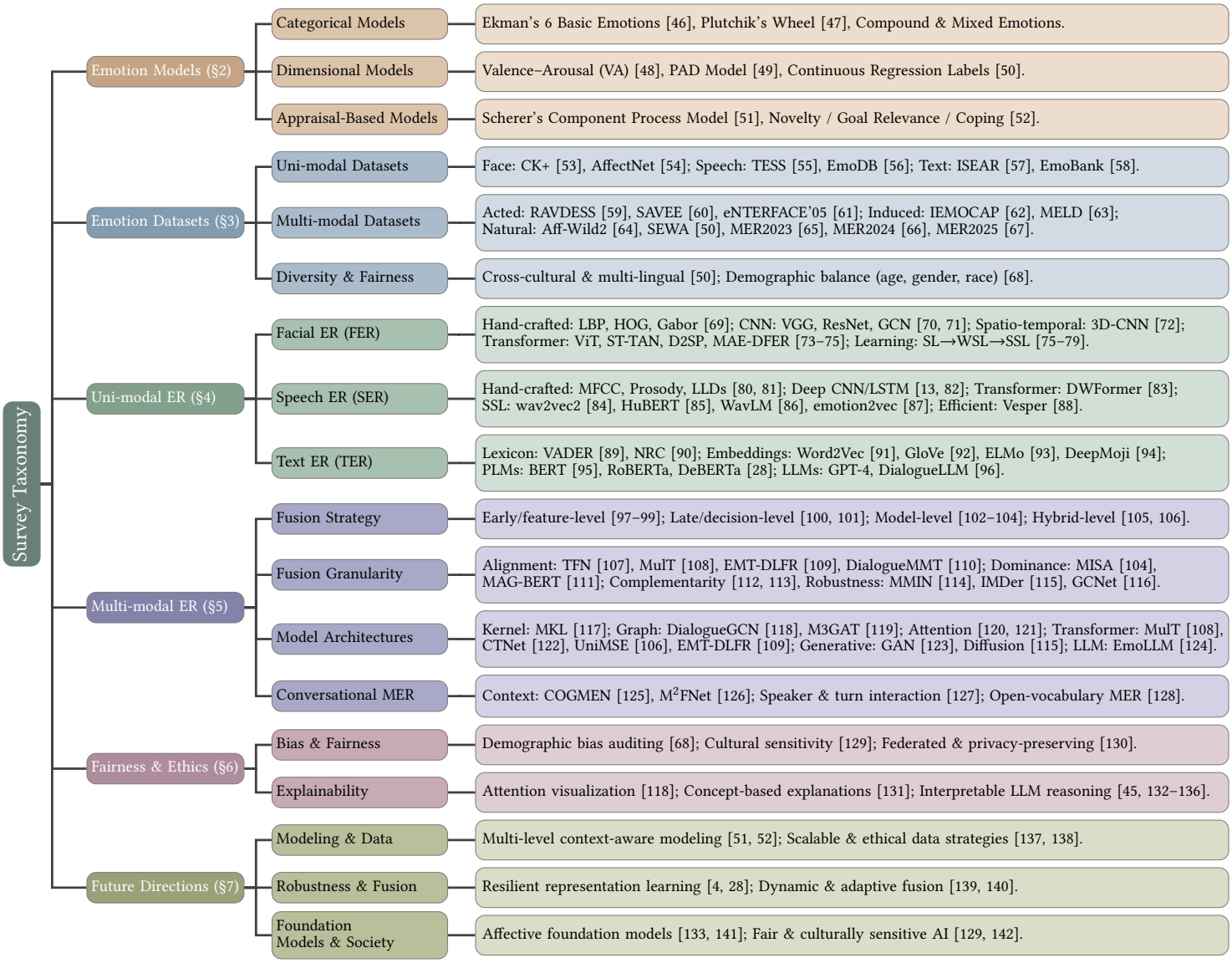


Fig. 4. Taxonomy of deep learning-based uni-modal and multi-modal emotion recognition.

explicitly linked to detailed analyses of model architectures, learning objectives, and datasets in Tables 4, 5, and 5, enabling transparent cross-study comparison.

- **Insights and Future Research Directions:** Based on the synthesized evidence, we distill key findings regarding emotion modeling choices, modality integration strategies, dataset construction, and evaluation practices. These insights are used to identify open challenges and to outline future research directions (Section 7), providing a roadmap toward more robust, generalizable, and ethically aligned emotion recognition systems.

By revisiting this field under the lens of our unified taxonomy and terminology, this survey consolidates disparate strands of emotion recognition research and establishes a coherent conceptual foundation for future work. The remainder of this paper is organized as follows. Section 2 introduces the theoretical foundations of emotion models and the modalities used to capture affective states. Section 3 surveys existing emotion recognition datasets, covering their collection protocols, annotation schemes, and key characteristics. Sections 4 and 5 systematically review uni-modal and multi-modal emotion recognition methods, respectively. Section 6 examines fairness considerations and ethical implications arising from the deployment of emotion recognition systems. Section 7 identifies open challenges and promising directions for future research. Finally, Section 8 concludes the review with a summary of key findings.

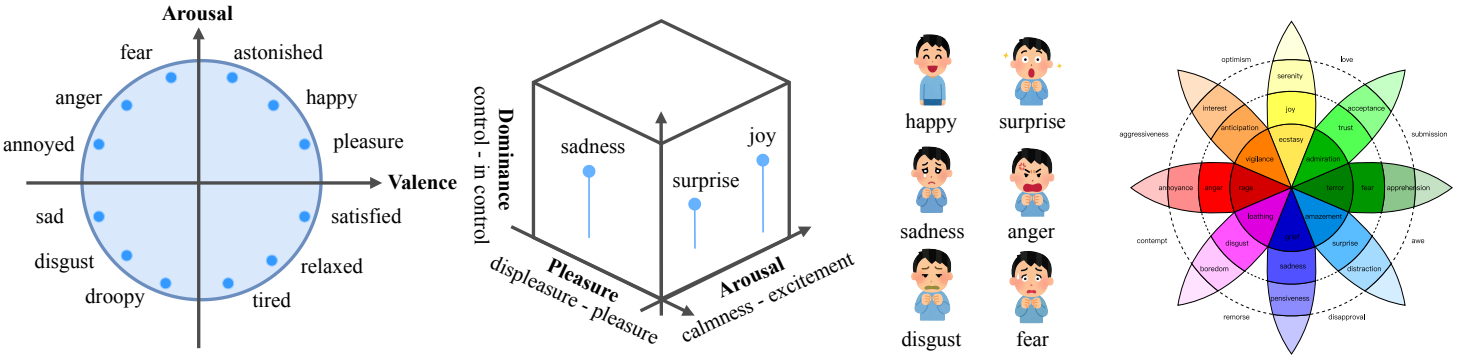
2 EMOTION MODELS AND MODALITIES

Understanding how emotions are modeled and expressed is essential for developing effective emotion recognition systems. Section 2.1 primarily discusses the different emotion models used in affective computing, and Section 2.2 elaborates on the various modalities through which emotions are expressed.

2.1 Emotion Models

Emotions are complex, multi-dimensional responses triggered by internal or external stimuli and accompanied by cognitive, physiological, and behavioral changes [143–145]. In affective computing, two dominant paradigms are used to model emotional states: categorical and dimensional models, with the latter developed to address limitations of the former. These theoretical models not only underpin how affective data are labeled and interpreted but also

directly influence dataset annotation protocols, model architectures, and evaluation criteria discussed later in Section 3, 4, 5, 6, 7. Understanding their conceptual distinctions is therefore crucial for interpreting cross-dataset comparisons and methodological trends throughout this survey.



(a) VA dimension emotion model (b) PAD dimension emotion model (c) Ekman's basic emotions (d) Plutchik's wheel emotions

Fig. 5. Comparison of Dimensional and Categorical Emotion Models: (a) Valence–Arousal (VA) model and (b) Pleasure–Arousal–Dominance (PAD) model for dimensional emotion representation; (c) Ekman's basic emotion categories and (d) Plutchik's emotion wheel for categorical emotion representation.

Categorical Emotion Models: Categorical emotion models represent the foundational approach to emotion representation, operating on the assumption that emotions can be classified into a discrete set of basic categories. Ekman's model [46] proposed six universally recognized emotions: happiness, sadness, anger, fear, disgust, and surprise (see Fig. 5(c)). These emotions are expressed through distinct facial muscle movements and behavior patterns and are widely used in image- and video-based emotion recognition systems. Another influential framework is Plutchik's wheel of emotions [47], which introduces eight bipolar primary emotions: joy vs. sadness, trust vs. disgust, fear vs. anger, and surprise vs. anticipation (see Fig. 5(d)). These primary emotions can be combined to generate more complex emotional states and further mapped by intensity levels.Categorical models offer significant practical advantages: they are intuitive for human annotators, straightforward to implement in classification tasks, and provide clear interpretability for users. Consequently, most publicly available datasets (e.g., FER, RAF-DB, CREMA-D) adopt categorical labels as their annotation scheme, facilitating model benchmarking and cross-modal comparison. However, their discrete nature and limited emotional vocabulary make them insufficient for representing the full spectrum and subtlety of real-world emotional experiences, particularly for capturing mixed emotions or gradual emotional transitions.

Dimensional Emotion Models: Dimensional emotion models go back to the works of Wilhelm Wundt, often referenced as "the father of psychology". He suggested the "tridimensional theory of emotion" as the first approach to model emotions in dimensions, lending inspiration to today's common approaches. It was based on the three axes pleasure-displeasure, excitement-tranquillization, and tension-relaxation. The models were developed to overcome the resolution limitations of categorical approaches by representing affective states in a continuous multi-dimensional space. Russell's circumplex model [48] defines emotions along only two of Wundt's original three axes—valence, indicating positivity or negativity, and arousal, indicating activation level (see Fig. 5(a)). Mehrabian and Russell later introduced as Wundt a third dimension, dominance, forming the Pleasure–Arousal–Dominance (PAD) model [49] (see Fig. 5(b)). The inclusion of dominance captures the perceived control or influence within emotional states, offering finer granularity than the basic valence–arousal framework. These continuous representations effectively model dynamic affective changes and are well-suited for regression-based learning, though they are often less intuitive and interpretable than discrete emotion categories.From a data perspective, many physiological and audio datasets (e.g., AMIGOS, DEAP, MSP-Podcast) adopt dimensional labels to describe fine-grained emotional variation over time, complementing categorical datasets introduced in Section 3. The coexistence of categorical and dimensional annotations thus shapes dataset design, learning objectives, and evaluation metrics across modalities, forming the conceptual foundation for consistent comparison of affective recognition systems throughout this survey.

2.2 Emotion Modalities

Human emotions manifest through behavior and physiological signals, both essential for emotion recognition [11, 146]. These modalities provide the empirical basis for datasets and models discussed later, linking theoretical emotion frameworks with measurable signals.

Behavior Modalities: Behavior cues, including facial expressions, speech, gestures, and text, serve as indicators of internal states [9, 17]. Facial expressions convey emotions but suffer from cultural bias and environmental interference [147, 148]. Speech encodes affect via prosody, further acoustics, and content, though speakers' further states and traits and noise complicate interpretation [13]. Text offers robustness and ease of processing but lacks paralinguistic richness [28, 149]. These modalities are popular due to accessibility and relevance in social interaction [30, 150]. Accordingly, many benchmark datasets employ behavior modalities under both categorical (e.g., discrete emotion labels) and dimensional (e.g., valence–arousal ratings) schemes, enabling comparative evaluation across representational models.

Physiological Modalities: Physiological signals provide objective markers of emotion, less prone to conscious regulation [151, 152]: EEG captures brain activity [153], ECG reflects heart rate variability [154], EMG tracks subtle muscular responses [155], and GSR detects skin conductivity changes [156]. Despite their precision, their reliance on sensors and complex processing limits broad deployment. Notably, physiological datasets frequently use dimensional models to encode continuous affective variation, complementing the discrete behavior datasets described later in Section 3.

Comparison of Modalities: Behavior signals are non-invasive, scalable, and socially meaningful but subject to ambiguity. Physiological signals provide objective, temporally precise, and physiologically grounded indicators of affective states. However, their reliance on contact-based sensors may reduce comfort and scalability while raising privacy concerns compared to behavior modalities. In practice, behavior cues, particularly audio, text, and video, are the dominant choices as they balance feasibility with expressiveness. This contrast in modality characteristics mirrors the categorical–dimensional divide: behavior modalities align naturally with discrete emotional expressions, whereas physiological modalities often capture continuous affective states, providing complementary perspectives for multi-modal emotion recognition.

Overall, emotion models and modalities jointly define the conceptual and empirical foundations of affective computing. Categorical and dimensional models offer complementary views of how emotions can be represented, while behavioral and physiological modalities provide the measurable signals through which emotions manifest. Their interplay determines how datasets are annotated, how models are designed, and how recognition performance is evaluated. Together, these frameworks establish a coherent bridge between psychological theory and computational implementation, setting the stage for the unimodal and multimodal emotion recognition methods discussed in subsequent sections.

Table 2. Summary of popular uni-modal and multi-modal emotion recognition datasets. Modality: A = Audio, T = Text, V = Visual.

Type	Dataset	Modality (A/V/T)	Emotion Labels	Samples
Uni-modal Emotion Recognition Datasets				
Face	CK+ [53]	V	Anger, disgust, fear, happiness, sadness, surprise, neutral, contempt	593 videos
	JAFFE [157]	V	Anger, disgust, fear, happiness, sadness, surprise, neutral	231 images
	Oulu-CASIA [158]	V	Anger, disgust, fear, happiness, sadness, surprise	2,880 videos
	SFEW 2.0 [159]	V	Anger, disgust, fear, happiness, sadness, surprise, neutral	1,766 images
	RAF-DB [160]	V	Anger, disgust, fear, happiness, sadness, surprise, neutral	29,672 images
	FER+ [161]	V	Anger, disgust, fear, happiness, sadness, surprise, neutral, contempt	35,887 images
	AffectNet [54]	V	Neutral, happy, sad, surprise, fear, disgust, anger, contempt	1,000,000 images
	EmoReact [162]	V	Curiosity, uncertainty, excitement, happiness, surprise, disgust, fear, frustration	1,102 videos
Speech	TESS [55]	A	Anger, disgust, fear, happiness, pleasant surprise, sadness, neutral	2,800 utterances
	EmoDB 2.0 [56]	A	Anger, boredom, disgust, fear, happiness, neutral, sadness	817 utterances
	MSP-Podcast [163]	A	Anger, contempt, disgust, fear, happiness, neutral, sadness, surprise	264,705 turns
Text	ISEAR [57]	T	Joy, fear, anger, sadness, disgust, shame, guilt	7,666 sentences
	EMOBANK [58]	T	Joy, anger, sad, fear, disgust, surprise	10,548 sentences
	SemEval-2018 [164]	T	Anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, neutral	22,000 sentences
Multi-modal Emotion Recognition Datasets				
Acted	eNTERFACE'05 [61]	A, V	Anger, disgust, fear, happiness, sadness, surprise	1,166 videos
	RAVDESS [59]	A, V	Calm, happy, sad, angry, fearful, surprise, disgust	7,356 videos
	CREMA-D [165]	A, V	Anger, disgust, fear, happiness, neutral, sadness	7,442 videos
	SAVEE [60]	A, V	Anger, disgust, fear, happiness, sadness, surprise, neutral	480 videos
Natural	AFEW [159]	A, V	Anger, disgust, fear, happiness, neutral, sadness, surprise	1,426 videos
	DFEW [166]	A, V	Anger, disgust, fear, happiness, neutral, sadness, surprise	16,372 videos
	FERV39k [167]	V	Anger, disgust, fear, happiness, neutral, sadness, surprise	38,935 videos
	MAFW [168]	A, V, T	11 single emotions and 32 compound emotions	10,045 videos
	CHEAVD [169]	A, V	Anger, happiness, sadness, worried, anxious, surprise, disgust, neutral	140 minutes video
	Aff-Wild2 [64]	V	Valence, arousal	558 videos
	SEWA [50]	A, V	Valence, arousal	2,000 minutes video
	AMIGOS [170]	A, V	Valence, arousal, dominance	40 videos
Induced	CMU-MOSI [171]	A, V, T	Continuous intensity score	3,702 videos
	CMU-MOSEI [137]	A, V, T	Happiness, sadness, anger, disgust, surprise, fear	223,500 videos
	MELD [63]	A, V, T	Anger, disgust, fear, joy, neutral, sadness, surprise	1,433 videos
	CH-SIMS [172]	A, V, T	Negative, weakly negative, neutral, weakly positive, positive	2,281 videos
	IEMOCAP [62]	A, V, T	Happiness, anger, sadness, frustration, neutral; valence, arousal, dominance	12.46 hour video
	RAMAS [173]	A, V	Anger, sadness, disgust, happiness, fear, surprise	7 hour video

3 EMOTION DATASETS

High-quality datasets are crucial for training effective automatic emotion recognition systems. The choice of data directly affects model performance, as poorly curated datasets may introduce noise and bias, while well-constructed datasets with consistent annotation enable robust, context-aware learning across modalities such as visual, audio, and text.

Single-modal datasets support specialized models by focusing on one modality. Facial datasets such as CK+ [53], and AffectNet [54] target expressions in images or video. Speech datasets like TESS [55], EmoDB [56], and MSP-Podcast [163] capture vocal affect, while text resources such as ISEAR [57], EmoBank [58], and SemEval-2018 Task 1 [164] provide labeled corpora for language-based recognition (Table 2).

Multi-modal datasets integrate two or more modalities and are typically categorized as *acted*, *induced*, or *natural* [9, 174, 175] (Table 2). Acted datasets (e.g., EmoDB [56], RAVDESS [59], SAVEE [176]) provide controlled, clean recordings but often exaggerate emotions. Induced datasets (e.g., IEMOCAP [62],

MELD [63], RECOLA [177]) elicit more authentic reactions through stimuli or tasks, capturing richer conversational dynamics. Natural(istic) datasets (e.g., Aff-Wild2 [64], CHEAVD [169], SEWA DB [50], AMIGOS [170]) collect spontaneous expressions in real-world settings, offering ecological validity but facing challenges in annotation, noise, and privacy.

Furthermore, as summarized in Table 2, a critical gap exists in the demographic representativeness of datasets. Early benchmarks like IEMOCAP and CK+ are notably homogeneous, whereas newer corpora like SEWA and MSP-Podcast begin to address cultural and age diversity, aligning with the growing demand for fair and inclusive affective computing (see Section 6).

Table 3. A Summary of Representative Datasets with Demographic Diversity Analysis.

Category	Dataset	Demographic Diversity & Insights	Diversity Level
Face	CK+ [53]	Lab-controlled; predominantly young Caucasian participants	Limited
	AffectNet [54]	Large-scale web data; global but lacks explicit balancing	Moderate
	FER+ [161]	Primarily internet images; exhibits Western-centric bias	Moderate
	Aff-Wild2 [64]	High diversity in race and age; captured in-the-wild	High
Speech	FACS [178]	High diversity in race and age	High
	RAVDESS [59]	Professional North American actors; limited racial variation	Limited
	MSP-Podcast [163]	Naturalistic speech with diverse accents and age groups	High
	CREMA-D [165]	Balanced ethnic representation across 48 actors	High
Text	Empathetic [179]	Crowdsourced dialogues; reflects native English norms	Limited
	GoEmotions [42]	Reddit-based; linguistically diverse but culturally Western	Moderate
Multimodal	IEMOCAP [62]	Very limited; only 10 actors from a similar demographic	Limited
	SEWA [50]	Cross-cultural; covers 6 nationalities and languages	High
	CMU-MOSEI [137]	High speaker diversity with 1000+ unique identities	High

*__High__: Datasets with explicit cross-cultural, multi-ethnic, or global coverage. __Moderate__: Datasets with large-scale or diverse samples but lacking demographic balancing or regional focus. __Limited__: Homogeneous datasets with participants from a single ethnic group, small group of actors, or restricted age range.

4 UNI-MODAL EMOTION RECOGNITION

This section surveys the development of emotion recognition across three major unimodal modalities, namely face, speech, and text, with a focus on the progressive shift from traditional hand-crafted feature engineering toward deep learning-based representation learning.

4.1 Facial Emotion Recognition

Facial emotion recognition (FER) has undergone a substantial methodological evolution, yet this progression is best understood not as a linear march of accuracy improvements, but as a recurring tension between representational richness and real-world robustness. Early methods relied on hand-crafted descriptors such as LBP, HOG, Gabor filters, and facial landmarks [69, 180, 181], which encoded appearance through fixed mathematical priors reflecting domain expertise about how emotional deformations manifest in local texture or geometry. The shift to CNN-based end-to-end learning [70, 182, 183] was therefore not merely a change in tooling but an epistemological one, marking a transition from theory-driven feature design to data-driven feature discovery. Notably, subsequent attention-enhanced architectures such as OAENet [184] and MA-Net [185], which reintroduce spatial selectivity over learned feature maps, can be interpreted as a partial rehabilitation of the locality priors that hand-crafted methods had always encoded, now learned rather than prescribed. GCN-based approaches [71, 186] similarly recover structural priors by explicitly modeling geometric relationships among facial landmarks. The trajectory from hand-crafted to deep features is thus less a clean break and more a dialectical refinement, each generation recovering in data-driven form what the previous one had abandoned.

A persistent and underappreciated limitation of static FER models is that they treat each frame as an independent observation, implicitly assuming that a single image encodes sufficient information to determine emotional state. This conflicts with psychophysical evidence that observers rely heavily on the temporal dynamics of facial muscle movement, specifically the onset, apex, and offset phases of an expression, to distinguish a genuine smile from a posed one, for example. The integration of sequential architectures [187, 188] and hybrid CNN–RNN pipelines such as SAANet [189] and MGLN [190] addressed this limitation, but their cascaded design decouples spatial and temporal learning, creating an information bottleneck between stages. Each generation of FER models has essentially inherited the inductive biases of the previous generation while patching its specific failure mode: unified spatio-temporal models such as DPCNet [72] and STACM [73] represent a more principled response by jointly modeling spatial, temporal, and channel dependencies through attention. More recently, Transformer-based approaches [74, 191] offer global temporal reasoning without the locality constraints of CNNs, though they typically initialize from ImageNet-pretrained weights, importing object-recognition biases into an affective domain, a representational mismatch that is rarely acknowledged but likely accounts for inconsistent gains across benchmarks.

Beyond architecture, FER has exhibited a clear paradigm shift in learning supervision, from fully supervised training toward weakly supervised, unsupervised, and self-supervised paradigms [75–79, 192–194]. This shift reflects a structural constraint that architectural innovation alone cannot resolve: emotion annotation is expensive, subjective, and culturally variable, and models trained on densely labeled laboratory datasets exhibit severe domain shift in the wild because the annotation protocol itself differs systematically across corpora. Self-supervised methods sidestep this by decoupling representation learning from label acquisition, deferring label assignment to a lightweight fine-tuning stage. The key implication is that the primary bottleneck in FER is not model capacity but annotation consistency; advances in label-efficient learning may therefore prove more impactful than continued architectural scaling.

Table 4. Unified comparison of facial, speech, and text emotion recognition models across architectures, loss designs, and reported performance.

Modality	Model	Input Format	Framework	Loss Function	Performance (Dataset)
Vision	C3D	Video	3D Conv	Softmax	Acc: 59.02% (AFEW) [195]
	I3D	Video	Inflated 3D	Softmax	Acc: 68.90% (GreSti) [196]
	SCE+TH-CNN	Video	SCE	Cross-Entropy	CCC Arousal: 65.6% (SEWA) [197]
	SlowFast	Video	Dual CNN	Softmax	WAR: 49.34% (FERV39K) [198]
	ViT-B/16/SAM	Video	Transformer	Cross-Entropy	Acc: 52.42% (FER-2013) [199]
	SL+SSL (B2)	Video	EfficientNet	CE+SL	Acc: 61.32% (AffectNet) [200]
	DTL-I-ResNet18	Video	3D ResNet	Softmax	Acc: 83.0% (FER2013) [201]
	ESTLNet	Video	CNN-LSTM	Cross-Entropy	Acc: 53.79% (AFEW) [72]
	D2SP	Video	Dual Purification	Cross-Entropy	WAR: 50.5% (FERV39k) [202]
	MAE-DFER	Video	Transformer	Cross-Entropy	WAR: 74.43% (DFEW), 52.07% (FERV39k) [75]
	SVFAP	Video	Transformer	Cross-Entropy	WAR: 74.27% (DFEW), 52.29% (FERV39k) [78]
Audio	Log Mel Spec.	MFCCs	2D CNN	Cross-Entropy	Acc: 68% (RAVDESS) [203]
	HuBERT	Raw audio	CNN+Transf.	Contrastive	WA: 79.58% (IEMOCAP) [204]
	Mockingjay	Raw audio	NPC	L1/MSE	Acc: 50.28% (IEMOCAP) [205]
	DeCoAR	Mel FBANK	SVM	L1/MSE	UAR: 71.93% (IEMOCAP) [206]
	NPC	Mel FBANK	Siamese CNN	Generative	Macro-F1: 30.4% (IEMOCAP) [207]
	APC	Mel spectrogram	RNN	Generative	Macro-F1: 31.6% (IEMOCAP) [207]
	VQ-APC	Mel spectrogram	VQ+RNN	Generative	Macro-F1: 31.2% (IEMOCAP) [207]
	Wav2vec	Raw audio	1D CNN	Contrastive	WA: 77.00% (IEMOCAP) [204]
	CPC	Raw audio	CNN+Transf.	Discriminative	Macro-F1: 28.5% (IEMOCAP) [207]
	Data2Vec	Raw audio	CNN+Transf.	Discriminative	Macro-F1: 33.9% (IEMOCAP) [207]
	emotion2vec	Raw audio	Online Distillation	Utterance & Frame-level	WA: 85.0% (RAVDESS) [87]
	WavLM	Raw audio	CNN+Transf.	Discriminative	Macro-F1: 33.6% (IEMOCAP) [207]
	Audio-Transf.	Spectrogram	Transformer	Cross-Entropy	Acc: 75.42% (EMO-DB) [208]
	Vesper	Raw audio	CNN+Transf.	MSE	WA: 54.2% (IEMOCAP) [88]
	SL-GEmo-CLAP	WavLM-large	CNN+Transf.	KL loss	WAR: 81.43% (IEMOCAP) [209]
	DTNet	Raw audio	CNN+Transf.	Cross-Entropy	UA: 74.8% (IEMOCAP) [210]
Text	Word2Vec	Text tokens	CBOW	Hierarchical Softmax	Macro-F1: 73.21% (Tweets) [211]
	GloVe	Text tokens	Co-occurrence matrix	Weighted Least Squares	Acc: 95.09% (Twitter) [212]
	ELMo	Context. vectors	BiLSTM	Cross-Entropy	Acc: 88.91% (Wikipedia) [213]
	Emoji2Vec	Emoji description	GloVe	Cosine Similarity	Acc: 54.5% (Google News) [214]
	SSWE	Word-sentiment	FFNN	Hinge Loss	Macro-F1: 84.98% (HL, MPQA) [211]
	BERT	Text token	Transformer	MLM + NSP	Acc: 70.09% (ISEAR) [215]
	RoBERTa	Text token	Transformer	Cross-Entropy	Acc: 74.31% (ISEAR) [215]
	XLNet	Permuted tokens	Transformer	Permuted LM Loss	Acc: 72.99% (ISEAR) [215]
	ALBERT	Text token	Transformer	Focal and KL loss	Acc: 73.86% (ISEAR) [215]
	DeBERTa-v3	Text token	Transformer	Cross-Entropy	F1: 66.2% (WRIME) [216]
	ChatGPT-4o	Text token	Transformer	N/A (Prompt-based)	F1: 52.7% (WRIME) [217]

(Continued on next page)

(Continued from previous page)

Modality	Model	Input Format	Framework	Loss Function	Performance (Dataset)
	DistilBERT	Text token	Transformer	MLM + Distillation	Acc: 66.93% (ISEAR) [215]
	COMET	Commonsense triple	Transformer	Cross-Entropy	W-Avg F1: 65.21% (MELD) [218]

4.2 Speech Emotion Recognition

Speech emotion recognition (SER) shares the broad arc of FER but the acoustic modality poses unique challenges that have driven distinct adaptations. Early SER relied on manually engineered prosodic, spectral, and voice quality features [80, 219], standardized through toolkits such as openEAR [220] and openSMILE [81] and combined with classical classifiers including HMMs, GMMs, SVMs, and Random Forests [221, 222].

A critical and often overlooked observation is that many of these features were originally designed for speech intelligibility and automatic speech recognition—and were partially repurposed for emotion; their coverage of affectively relevant acoustic phenomena such as creaky voice, breathiness, and micro-prosodic perturbations is therefore often incomplete by design. End-to-end learning from raw waveforms [82] demonstrated that data-driven features could capture affective cues invisible to pre-specified filter banks, and spectrogram-based CNN and LSTM models [223–225] became dominant supervised baselines. Yet this progress exposed what might be termed a feature design paradox: hand-crafted features encode the wrong priors for emotion, but learned features require large quantities of emotion-labeled data that SER datasets, which typically comprise only a few hours of acted speech, simply cannot provide, thereby constraining generalization in cross-corpus and low-resource settings [13].

Recent SER research has seen Transformer-based architectures [83, 226–228] improve contextual modeling of speech emotion, while parameter-efficient fine-tuning strategies [229] and compact pretrained models [88] target deployment under resource constraints. More substantively, contrastive and adversarial frameworks such as STAA-Net [230] and GEmo-CLAP [209] begin to address demographic bias, which remains a serious concern when SER systems behave inequitably across speakers of different genders, ages, or dialects. The community currently lacks standardized evaluation protocols for demographic fairness in SER, and this gap between laboratory benchmarks and real-world deployment requirements remains an under-addressed structural challenge.

Building on supervised learning, self-supervised learning (SSL) resolves the feature design paradox structurally rather than incrementally. By pre-training on thousands of hours of unlabeled speech, models such as wav2vec [84], HuBERT [85], WavLM [86], and data2vec [231] learn representations that remain robust across different speakers and recording conditions, because the pre-training distribution is far broader than any emotion corpus. The consistent cross-corpus gains documented by Wagner et al. [232] and Naini et al. [233] are a direct consequence of this property: SSL models are less prone to learning spurious patterns tied to specific datasets, such as confounding recording loudness with anger, than models trained exclusively on small, biased emotion datasets. Critically, emotion-aware variants such as emotion2vec [87] demonstrate that injecting affective objectives during pretraining, not only during fine-tuning, produces representations that are simultaneously transferable and emotion-discriminative. This finding implies that the optimal inductive bias for SER lies somewhere between generic acoustic pretraining and fully supervised emotion training, a middle ground that the field is only beginning to explore systematically.

4.3 Text Emotion Recognition

Textual emotion recognition (TER) has progressed from lexicon-based methods to large-scale pretrained language models, but this trajectory conceals a recurring tension between representational expressiveness and the availability of labeled data. Early approaches combined affective lexicons such as VADER [89] and the NRC Emotion Lexicon [90] with classical classifiers and surface-level features including n-grams [234], POS tags [235], and syntactic patterns [236]. These methods were interpretable and computationally efficient, but they systematically failed on conversational and social media text because phenomena such as irony and sarcasm can reverse the apparent meaning of individual words in ways that fixed lexical mappings cannot capture. Static embeddings [91, 92, 237] and emotion-aware variants such as Emo2Vec [238] and DeepMoji [94] reduced lexical sparsity, and CNN- and LSTM-based models [239–241] provided stronger sequence-level representations. Yet these supervised architectures remained heavily dependent on large annotated corpora that are expensive and subjective to construct, exposing a structural mismatch between model capacity and data availability that lexical and neural methods alike failed to resolve.

Transformer-based pretraining fundamentally changed this dynamic by decoupling representation learning from emotion-labeled data. Contextualized models such as ELMo [93] and BERT [95], and their emotion-specific fine-tuned variants including EmoDet2 [242], EmotionX-IDEA [243], and RoBERTa-based classifiers [244], consistently outperformed prior architectures by capturing word sense ambiguity and long-range contextual dependencies. Subsequent work incorporated commonsense knowledge, sarcasm-aware attention, and parameter-efficient fine-tuning to improve robustness in cross-domain settings [28]. However, a critical observation is that these gains are largely attributable to better general language understanding rather than to any deeper modeling of emotion itself, meaning that pretrained models inherit the biases of their pretraining corpora, which are rarely balanced for affective content or demographic diversity.

More recently, large language models (LLMs) have emerged as a paradigm-shifting force in TER, fundamentally extending the field beyond conventional supervised learning frameworks. By leveraging large-scale pretraining and instruction-based inference, LLMs exhibit strong zero-shot and few-shot emotion recognition capabilities, substantially reducing reliance on task-specific labeled datasets and enabling emotion understanding to be framed as a general language reasoning problem. Representative studies have shown that generalized large models possess emergent emotion reasoning abilities across domains, languages, and emotion taxonomies, marking a departure from narrowly trained classifiers [245]. Building upon this foundation, emotion-aware and context-enhanced LLMs, such as DialogueLLM, explicitly integrate conversational context and emotion knowledge to model emotion dynamics in multi-turn dialogues [96], while data-centric approaches like AUGESC employ LLMs to augment emotionally rich conversational data, alleviating data scarcity in emotional support scenarios [246]. Beyond text-only settings, recent advances demonstrate that LLMs can be extended toward generalized and multi-modal emotion recognition, incorporating vocal nuances or visual signals to infer affective states without task-specific training [247, 248]. Despite their strong generalization and reasoning capabilities, practical deployment challenges remain due to computational cost and latency, motivating ongoing research into lightweight variants and parameter-efficient adaptation strategies for real-time and resource-constrained emotion-aware applications.

4.4 Summary of Uni-modal Emotion Recognition

Reviewing FER, SER, and TER together reveals cross-cutting patterns that transcend any individual modality. Most fundamentally, all three fields have converged on the same structural response to the annotation bottleneck, namely decoupling representation learning from label acquisition through self-supervised or weakly supervised pretraining, a convergence that suggests unified cross-modal pretraining frameworks could address shared annotation scarcity more effectively than modality-specific solutions developed in isolation. Beyond this commonality, each modality exhibits characteristic failure conditions, with FER degrading under occlusion and pose variation, SER under speaker variability and recording noise, and TER under sarcasm and implicit affect, and these failure conditions are largely independent of one another. The benefit of multi-modal fusion is therefore largest precisely in the hard cases where any single modality fails, a benefit that single-modality benchmarks systematically underestimate by design. Finally, across all three modalities, the gap between benchmark accuracy and performance in naturalistic settings has not narrowed commensurate with architectural advances, pointing to distribution shift and annotation inconsistency as the primary impediments to real-world deployment rather than insufficient model capacity. These shared limitations motivate the multi-modal fusion approaches examined in the following section, which seek to exploit cross-modal complementarity to achieve robustness that no single stream of evidence can provide alone.

5 MULTI-MODAL EMOTION RECOGNITION

Human emotions are inherently expressed in a multi-modal manner, and effectively integrating heterogeneous cues from multiple modalities is crucial for robust emotion recognition in conversations [249]. Accordingly, this section reorganizes multi-modal emotion recognition methods along three key aspects: *fusion strategy*, *fusion granularity*, and *model architecture*.

5.1 Fusion Strategy

Multi-modal emotion recognition requires integrating signals from heterogeneous sources, and the choice of fusion strategy fundamentally determines what kind of cross-modal reasoning is possible. Conventional fusion approaches fall into four broad categories: *feature-level*, *decision-level*, *model-level*, and *hybrid fusion* [3, 250]. Rather than being interchangeable design choices, these strategies embody different assumptions about when and how modalities should interact, and their relative strengths reveal a consistent trade-off between cross-modal expressiveness and robustness to noisy or missing inputs.

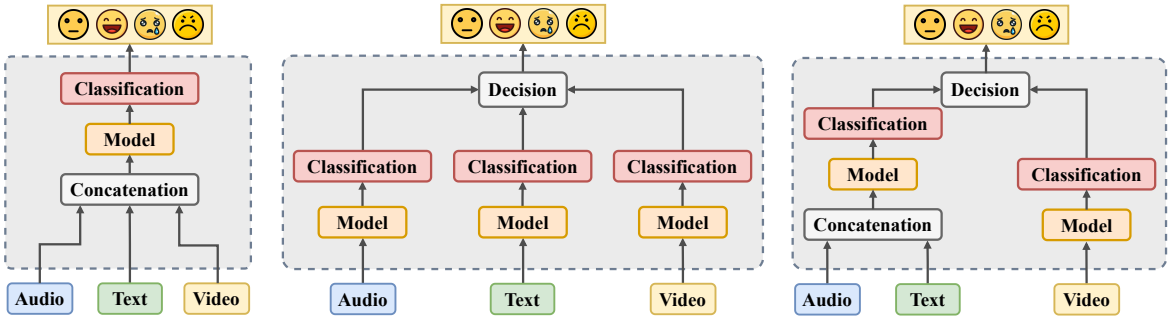


Fig. 6. (a) Early fusion, (b) late fusion, and (c) hybrid fusion methods for fusion strategy in multi-modal emotion recognition.

Feature-level fusion integrates multi-modal representations early in the pipeline (see Fig. 6(a)), enabling joint learning of low-level cross-modal correlations through mechanisms such as contrastive alignment [97], attention-based selection [98], and cross-attention in Transformer frameworks [99, 251]. The underlying assumption is that emotional semantics emerge from the co-occurrence of fine-grained cues across modalities, a reasonable prior for well-aligned, laboratory-quality recordings. However, this assumption breaks down under temporal asynchrony or noisy inputs, precisely the conditions that characterize real-world deployment, because early fusion leaves the model no mechanism to discount an unreliable modality once their representations have been entangled. Decision-level fusion decouples feature learning entirely (see Fig. 6(b)), aggregating modality-specific predictions through confidence-aware weighting [100] or graph-based consistency modeling [101]. This modularity provides natural robustness to missing or corrupted modalities, but it sacrifices the deep cross-modal interaction that distinguishes genuinely multi-modal reasoning from a simple ensemble of independent classifiers.

Model-level fusion occupies a principled middle ground, where modality-specific encoders are retained but cross-modal interactions are introduced at intermediate semantic layers through shared memory [102], low-rank tensor decomposition [103], or disentangled modality-invariant and modality-specific representations [104, 252, 253]. By deferring fusion to a stage where modalities have already been partially abstracted, model-level approaches reduce sensitivity to low-level temporal misalignment while preserving richer inter-modal dependencies than decision-level fusion permits. Hybrid fusion extends this logic further by introducing cross-modal interactions at multiple stages simultaneously (see Fig. 6(c)), ranging from directional cross-attention [105] and gated injection of acoustic and visual signals into pretrained language models [254], to lightweight adapter-based integration [106], progressively exchanging information across modalities rather than committing to a single fusion point. The insight that emerges across all four strategies is that no single fusion design is universally optimal: the best choice depends on whether the priority is expressive cross-modal interaction, robustness to real-world noise, or computational efficiency, and the field is only beginning to develop principled criteria for making this choice in a given deployment context.

Table 5. Comprehensive summary of recent fusion strategy and fusion granularity methods for multi-modal emotion recognition (MER), covering fusion types, techniques, advantages, and disadvantages.

Fusion Method	Specific Techniques	Advantages	Disadvantages
Fusion Strategy			
Early Fusion	Concatenation [255] Contrastive Learning [256] Transf.-based [126] Attention-based [112]	• Simple, efficient • Early interaction captured • Improves modality alignment • Reduces redundancy	• Can lose modality-specific details • Temporal misalignment is difficult to handle • May suffer from information overload
Late Fusion	Decision Voting [257] Weighted Average [258] Ensemble [102]	• Preserves intra-modal information • Ideal for real-time applications • Easy to implement • Increases model robustness	• Computationally expensive • Requires multiple classifiers • Long inference time due to multiple stages
Hybrid Fusion	Multi-stage [259] Early-Late [105] Intermediate [260]	• Captures intra- and inter-modal dependencies • Robust to noisy conditions • Flexible for real-world applications	• Computationally heavy • Hard to balance modality contributions • Complex integration process
Fusion Granularity			
Modality Alignment	Coarse-grained utterance-level [107] Structure-aware [118] Semantic alignment [261] Distribution-aware [110]	• Preserves modality-specific features • Simple and efficient utterance-level alignment • Does not require fine-grained correspondence supervision	• Alignment remains coarse and implicit • Limited fine-grained cross-modal interaction • Sensitive to temporal and semantic misalignment
Modality Dominance	Coarse-grained fusion [262] Memory-based fusion [102] Adaptive mechanism [263]	• Simple fusion without explicit modality balancing • Allows dominant modalities to exploit strong semantic cues • Easy to optimize with standard classifiers	• Prone to text-centric dominance • Suppresses weaker audio and visual cues • Degrades robustness when dominant modality is noisy or biased
Modality Complementarity	Attention-driven fusion [112] Model-level fusion [113] Semantics enhancement [264]	• Encourages synergistic use of heterogeneous modality cues • Highlights non-redundant emotional evidence • Preserves fine-grained affective signals across modalities	• Depends on accurate attention or interaction design • Susceptible to noise amplification from unreliable modalities • Increased modeling complexity and computational cost
Modality Robustness	Reconstruction fusion [114] Prompt learning [265] Diffusion-based [115]	• Maintains performance under missing or incomplete modalities • Enables graceful degradation • Explicitly models modality availability or recovery	• Relies on accurate reconstruction or conditioning signals • Error accumulation may occur in severe missing-modality scenarios • Increased training and inference complexity

5.2 Fusion Granularity

Fusion granularity determines the resolution at which modalities interact, and its design is governed by four intertwined challenges: *modality alignment*, *modality dominance*, *modality complementarity*, and *modality robustness* (see Fig. 7). These are not independent engineering problems but facets of a single underlying tension, namely that finer granularity enables richer cross-modal interaction but amplifies sensitivity to noise, misalignment, and missing data, while coarser granularity sacrifices expressiveness for stability. The evolution of fusion granularity in MER reflects a gradual shift from assumption-driven global pooling toward implicitly learned, structure-aware, and uncertainty-aware interaction mechanisms, driven by the progressive recognition that each of these four challenges demands fine-grained, principled treatment rather than heuristic patching [4, 12, 36, 139, 266–271].

5.2.1 Modality Alignment. Modality alignment refers to establishing meaningful correspondences between sub-components of different modalities so that cross-modal representations convey compatible semantics at appropriate temporal and structural scales [272, 273]. In emotion recognition, this is

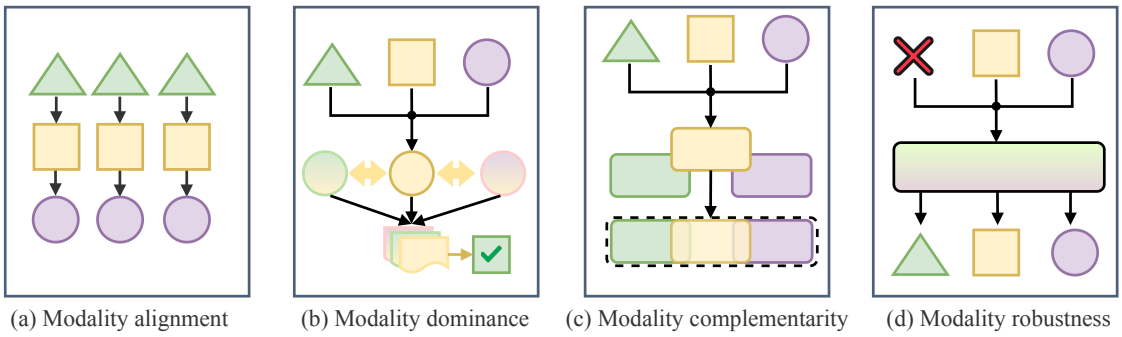


Fig. 7. (a) Modality alignment, (b) modality dominance, (c) modality complementarity, and (d) modality robustness as four core challenges of fusion granularity in multi-modal emotion recognition.

particularly difficult because affective cues are heterogeneous and weakly synchronized: acoustic prosody evolves continuously, lexical tokens discretize semantic meaning, and facial expressions may precede or lag spoken language. Although explicit alignment, which directly maps speech frames to transcript tokens or facial segments to linguistic units, offers precision, it requires fine-grained correspondence annotations that emotion datasets rarely provide, and affective expressions are often diffuse and context-dependent rather than localized to precise tokens or frames. As a result, most MER systems adopt implicit alignment, where cross-modal synchronization emerges as a latent process guided by downstream emotion prediction rather than by direct supervision [272]. The critical insight here is that this pragmatic choice carries a hidden cost: models that never receive explicit alignment supervision may learn spurious cross-modal associations that happen to correlate with emotion labels in controlled datasets but fail to generalize when modalities are asynchronous or one is degraded.

The evolution of implicit alignment in MER reflects a gradual shift from coarse global assumptions toward finer-grained, structure-aware mechanisms. Early utterance-level fusion [274–276] assumed modalities were roughly synchronized, an assumption that is computationally convenient but fragile in practice. Subsequent models introduced learnable cross-modal interactions to allow alignment to emerge from the data: TFN [107] captures correlated affective cues through multiplicative cross-modal interactions, RMFN [277] coordinates modalities over time via recurrent memory, and MulT [108] establishes cross-modal attention as a flexible template for aligning unaligned sequences. As emotion recognition moved into conversational settings, alignment requirements became more demanding, requiring consistency not only within utterances but across dialogue turns as emotional states evolve with speaker interaction. Structure-aware models such as MMGCN [118] and M^{2}FNet [126] address this by modeling inter-utterance dependencies explicitly, while more recent approaches such as DialogueMMT [110] and Ugncl fusion [278] further separate modality representations into shared and private components to control domain-specific drift during alignment. Taken together, these developments suggest that robust alignment in real-world emotion recognition cannot be treated as a preprocessing step or a byproduct of fusion, but must be treated as a first-class modeling objective in its own right.

5.2.2 Modality Dominance. Modality dominance refers to the tendency for one modality, most commonly text, to disproportionately influence the fused representation due to its richer semantic content, higher feature dimensionality, or stronger supervisory signals. In practice, this imbalance causes models to exploit shortcut linguistic cues while suppressing contributions from audio and visual modalities, ultimately degrading generalization when textual input is noisy, biased, or unavailable [3, 18]. The deeper issue is that modality dominance is not a transient training artifact that can be corrected through better optimization, but a structural consequence of fusion design itself: systems that concatenate independently encoded modality representations [102, 262, 279] implicitly permit classifiers to ignore weaker acoustic and visual streams whenever the textual signal is sufficiently discriminative for the training labels. This means that a model can achieve strong benchmark performance while remaining functionally unimodal, a failure mode that standard evaluation protocols based on balanced, clean, laboratory-collected datasets are poorly equipped to detect. Dominance that emerges at early fusion stages propagates through subsequent layers and becomes increasingly difficult to reverse, making it a problem that must be addressed by design rather than corrected after the fact [272].

The field has responded with progressively more principled strategies, and tracing their evolution reveals an important shift in how the problem is conceptualized. Early approaches treated dominance as a representation problem and addressed it through factorization and disentanglement [104, 280–282], decomposing each modality into shared and private subspaces to prevent any single modality from saturating the joint representation. A complementary direction used pretrained language models as controlled fusion backbones, where acoustic and visual signals modulate textual representations through constrained gating rather than unconstrained feature injection, as in MAG-style adaptation [111], limiting the degree to which text can monopolize the fused output. Context-aware mechanisms further extended this logic to conversational settings, where dominance patterns shift dynamically across dialogue turns: TELME [263] uses a teacher–student framework to penalize text-only shortcuts, CSS [283] adaptively re-weights modality contributions at each turn, and AcFormer [284] demonstrates that dominance control must be explicitly maintained under efficiency constraints or models revert to text-centric behavior. More recent work reframes dominance as an inference problem rather than a representation problem: evidence-centric frameworks such as ECERC [285] and ESED [286] restructure the decision process around multi-modal evidence graphs and uncertainty-aware reasoning, so that the model must actively justify its predictions using all available modalities rather than defaulting to the most convenient one. Similarly, AmBER^2 [287] further models modality disagreement by incorporating rater-variability into training objective and adaptively penalising conflicting modalities. Causal approaches such as CIDer [288] and reliability-aware models such as TiCAL [289] further disentangle genuine affective factors from spurious modality-specific correlations, dynamically down-weighting dominant but unreliable signals at inference time. The central insight that emerges from this progression is that solving modality dominance requires intervening at the level of how models reason about evidence, not merely at the level of how features are combined, and that evaluation protocols must be redesigned to reward genuinely multimodal reasoning rather than text-centric benchmark performance.

5.2.3 Modality Complementarity. Modality complementarity in MER concerns whether fusion can convert heterogeneous emotional evidence from text, audio, and vision into a genuinely synergistic representation, rather than a redundant aggregation that merely repeats the dominant modality's signal. The distinction matters because a model that appears to perform multi-modal fusion may in practice be learning to replicate what text alone already provides, with audio and visual streams contributing little beyond noise. A key insight is that complementarity is not an inherent property of the modalities themselves but an emergent property of how fusion is designed: coarse utterance-level pooling tends to dilute modality-specific cues such as prosodic contours or micro-expressions, whereas fine-grained token-frame-region interaction preserves precisely the non-overlapping affective evidence that makes multi-modal reasoning more powerful than any single modality alone. This means that improving complementarity is fundamentally a question of fusion granularity rather than simply adding more modalities.

Early work operationalized complementarity by designing fusion modules that explicitly separate shared from modality-specific information and selectively emphasize the most informative cues. Liu et al. [290] explicitly model both complementary information and modality importance to avoid treating modalities as interchangeable signals, while attention-driven fusion demonstrates that complementarity emerges most reliably when interaction is selective rather than uniform, since attention can highlight cross-modal segments that provide non-overlapping affective evidence [112]. The subsequent shift toward hierarchical and local-global interaction reflected a deeper understanding of why single-stage fusion is insufficient: emotional evidence is distributed across both modalities and conversational turns, so capturing complementarity requires multi-scale interaction that couples local emotional cues with broader dialogue context. Hierarchical cross-modal spatial fusion [113] organizes interaction across spatial and temporal resolutions so that subtle local cues are not overwritten by global pooling, while HiMul-LGG [291] propagates complementary cues between utterance-level dynamics and dialogue-level structure through local-global graph modeling. More recent work makes complementarity increasingly structured and transferable: dynamic graph neural ODE modeling [292] treats cross-modal interaction as a continuous-time dependency evolution process, preserving complementary cues that appear at different temporal granularities; graph-spectrum analysis [293] provides diagnostic tools for understanding when graph-based fusion amplifies or suppresses complementary information; and semantics-aware frameworks such as CIME [264] encourage complementary evidence to be assembled into emotion-consistent meaning rather than mere feature overlap. DEEMO [294] further demonstrates that removing identity-correlated shortcuts implicitly forces models to rely on genuinely complementary affective cues, improving transfer across subjects and domains. Taken together, these developments point to a broader principle: achieving robust complementarity requires not only finer fusion granularity but also explicit mechanisms that prevent models from collapsing back onto the easiest available signal, whether that signal comes from a dominant modality, a dataset-specific shortcut, or identity-correlated features that happen to correlate with emotion labels in controlled settings.

5.2.4 Modality Robustness. Modality robustness in MER refers to the ability of a fusion system to maintain reliable performance when one or more input streams are corrupted, incomplete, or entirely absent. In practical deployment, visual signals may be occluded, audio streams corrupted by noise or silence, and textual inputs missing due to ASR failures, latency, or privacy constraints. A critical insight that early empirical work established is that robustness is not an inherent byproduct of multimodality: naive fusion strategies that perform well when all modalities are present often degrade sharply when one is missing at inference time [114, 295], because the fusion operator has implicitly learned to rely on whichever modality provided the strongest training signal. This means that a system can appear robust during evaluation on complete data while remaining brittle in exactly the deployment conditions where multi-modal sensing is most likely to fail. MissModal [296] makes this failure mode explicit by framing missing modality as a systematic train-test mismatch, showing that fusion mechanisms over-committed to a dominant modality can collapse catastrophically when that modality becomes unavailable. The implication is that robustness must be engineered into the fusion process from the outset rather than treated as a property that emerges naturally from combining multiple streams.

As the field progressed, the strategy shifted from tolerating missing modalities to actively recovering them. Rather than adapting fusion to incomplete inputs, models began to infer missing modality representations from available ones. The Missing Modality Imagination Network [114] frames this as a conditional reconstruction problem that jointly optimizes imagination and emotion prediction, while MMIN introduces modality-invariant feature learning to ensure that reconstructed representations preserve task-relevant affective content rather than superficial statistical correlations. Translation-based approaches reinterpret missingness as a cross-modal mapping problem, projecting available modalities into the representation space of absent ones to stabilize downstream inference [297]. More recent work extends this further by making modality availability an explicit conditioning signal rather than an implicit assumption: TATE [298] encodes missingness patterns via modality tags, and multi-modal prompt learning [265] conditions prediction on which modalities are present, improving generalization across arbitrary modality subsets. The most recent advances couple recovery with uncertainty modeling through generative frameworks: IMDer [115] treats incomplete multi-modal inputs as a diffusion-based recovery problem, and RRMER-DT [299] integrates diffusion-based recovery with Transformer fusion in conversational settings. Taken together, these developments reflect a broader conceptual shift in how the field understands robustness: from a passive property of fusion architectures to an active modeling objective that requires explicitly reasoning about what is missing, why it is missing, and how reliably it can be recovered from the remaining streams.

5.3 Model Architectures

In this section, modern multi-modal fusion architectures in MER can be broadly classified into *seven categories*: (1) *kernel-based architectures*, (2) *graphical architectures*, (3) *neural network–based architectures*, (4) *Transformer-based architectures*, (5) *attention-based architectures*, (6) *generative-based architectures*, and (7) *large language model-based architectures*. (see Fig. 8,Fig. 9, and Fig. 10).

5.3.1 Kernel-based Architectures. Multiple kernel learning (MKL) established the first principled framework for multi-modal fusion in MER by associating each modality with its own similarity function and learning an optimal weighted combination within a unified convex objective [117]. Its enduring conceptual contribution is the explicit recognition that heterogeneous affective signals, including facial appearance, speech prosody, text, and physiological signals, cannot be meaningfully compared through a single shared kernel because their statistical geometries differ fundamentally. Hybrid extensions that combine MKL with deep convolutional encoders [5, 274, 300] preserved this principle of modality-specific similarity modeling while replacing handcrafted kernels with learned representations. MKL was ultimately superseded by end-to-end neural fusion due to its scalability limitations and dependence on support vectors, but its foundational principles of modality-specific encoding, weighted complementary fusion, and explicit similarity modeling continue to inform the design of modern architectures, often implicitly.

5.3.2 Graphical Architectures. Graph-based architectures address a limitation that neither kernel nor flat neural models can resolve: in conversational MER, emotional meaning is not localized within individual utterances but emerges from relational structures spanning speakers, turns, and modalities.

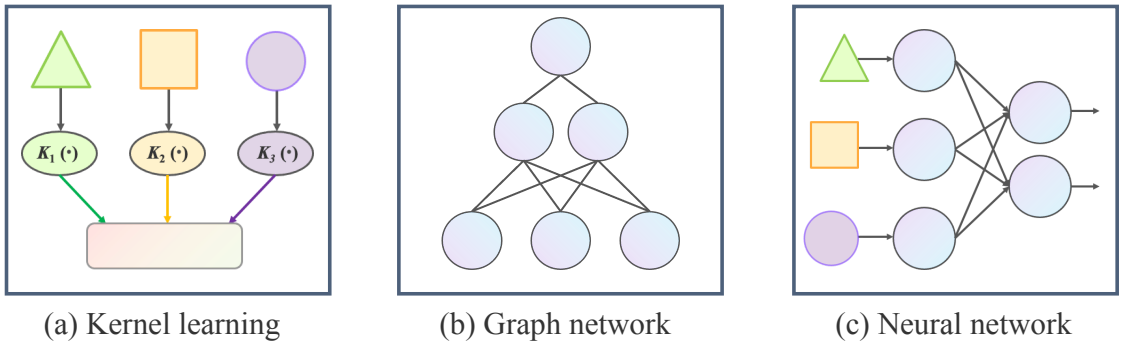


Fig. 8. (a) kernel-based architectures, (b) graphical architectures, and (c) neural network-based architectures in multi-modal emotion recognition.

The critical insight motivating this paradigm is that relational structure is itself affective information. Whether an utterance expresses frustration depends not only on its acoustic and lexical content but on who said it, to whom, and what was said before. Models such as COGMEN [125] and M3GAT [119] operationalize this by treating dialogue as a heterogeneous graph where nodes represent utterances and edges encode speaker, temporal, and cross-modal dependencies. More recent extensions further differentiate utterance-, speaker-, and modality-level relational structures [301], incorporate persona and emotion shift detection [127], and align modality-specific with modality-invariant representations through graph-based distillation [302]. Taken together, graphical models represent the most structurally explicit approach to conversational MER, and their continued development reflects a recognition that scalable emotion understanding in dialogue requires reasoning over relational context, not merely aggregating per-utterance features.

5.3.3 Neural Network-based Architectures. Early neural MER models established a template that has proven remarkably persistent: separate uni-modal encoders, typically CNNs or LSTMs, followed by joint fusion in shared hidden layers. Foundational work by Wöllmer et al. [303] and Nicolaou et al. [304] showed that temporal emotion dynamics could be modeled through recurrent fusion without hand-crafted pipelines, while Tzirakis et al. [262] demonstrated fully end-to-end audio-visual learning for continuous emotion recognition. Ensemble CNNs further improved generalization under limited labeled data [305], and semi-supervised approaches [306] extended this to sparse annotation settings. Yet, a structural limitation runs through all of these designs: the fusion bottleneck resides in the classifier, not the encoder. When modality representations are simply concatenated before a shared output layer, the classifier is free to down-weight or ignore weaker modalities whenever the dominant one already minimizes training loss, and nothing in the objective penalizes this shortcut. This is not a failure of specific models but an inherent consequence of the concatenate-then-classify paradigm.

Later work addressed this through more principled fusion objectives and cross-subject generalization strategies. MIST [307] combines DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial recognition, and 3D-CNN for motion analysis within a multi-modal framework that improves robustness under limited data, while DISD-Net [308] incorporates dynamic interaction and self-distillation for cross-subject emotion recognition. These advances reflect a broader recognition that generalization failures in neural MER are not primarily a capacity problem but a distribution mismatch problem: models trained on acted or laboratory data fail in the wild because their fusion weights are calibrated to the statistics of the training corpus rather than to the underlying affective signal. Interpretability remains an unresolved gap across this entire paradigm, as it is generally unclear which modalities or temporal segments a neural MER model relies on for a given prediction, limiting deployment in applications where accountability is required.

Table 6. Comprehensive summary of model architectures for multi-modal emotion recognition, covering fusion types, techniques, advantages, and disadvantages.

Model Architectures	Specific Techniques	Advantages	Disadvantages
Kernel-based Fusion	Modality-specific kernels fusion [117] MKL-based fusion [5]	• Supports heterogeneous modalities via modality-specific kernels • Enables principled weighted fusion with convex optimization • Effective for modeling complementary similarity patterns	• Limited scalability with large datasets • High computational and inference cost • Difficult to integrate into end-to-end deep architectures
Graphical-based Fusion	Capsule graph convolutional fusion [309] Attention-based graph structure [119] Hierarchical heterogeneous graph [301] Decoupled distillation graph [302]	• Explicitly models structured dependencies across modalities and context • Captures complementary cues via relational reasoning • Well-suited for conversational and context-aware MER	• Performance sensitive to graph construction quality • Increased architectural and computational complexity • Difficult to scale to long or densely connected dialogues

(Continued on next page)

(Continued from previous page)

Fusion Method	Specific Techniques	Advantages	Disadvantages
Neural Networks-based Fusion	RNNs and LSTMs structure [304] CNN structure [305] Semi-supervised learning [308]	• Learns complex non-linear cross-modal interactions • Effective for temporal and continuous emotion modeling • Scales well with large training datasets	• Requires substantial labeled data • Limited interpretability of fusion decisions • Performance sensitive to data quality and domain shift
Attention-Based Fusion	Self-attention [120] Cross-attention [121] Spatial-attention [310] Temporal-attention [311] Contextual-attention [312]	• Selectively highlights emotionally salient cues • Effectively handles modality heterogeneity and asynchrony • Supports fine-grained spatial, temporal, and contextual fusion	• Sensitive to attention weight estimation • May amplify spurious or noisy cues • Additional design complexity across attention types
Transformer-Based Fusion	Self-supervised Transformer [313] Pre-trained [314, 315] Unified Transformer [316]	• Models long-range temporal and cross-modal dependencies • Flexible attention-based fusion for asynchronous modalities • Benefits from large-scale pretraining and self-supervision	• Computationally expensive for long sequences • Performance sensitive to attention design and data bias • Limited interpretability of learned attention patterns
Generative-Based Fusion	GAN-based fusion [123] Diffusion-based fusion [115] Autoencoder-based fusion [317]	• Models underlying data distributions for robust fusion • Enables modality reconstruction and data augmentation • Improves robustness under missing or noisy modalities	• Training and inference are computationally intensive • Generation quality strongly affects downstream fusion • Optimization is often unstable and hard to tune
LLM-Based Fusion	LLMs-based fusion [124] Fine-tuning fusion [318] Prompt-based learning [319]	• Leverages strong semantic reasoning and world knowledge • Supports instruction tuning and prompt-based fusion • Enables zero-shot and open-vocabulary emotion recognition	• High computational and memory cost • Sensitive to prompt design and data bias • Limited control over fine-grained non-textual cues

5.3.4 Attention-based Architectures. Attention mechanisms in MER span a spectrum of interaction granularity that is often collapsed into a single category in the literature, obscuring important functional distinctions (see Fig. 9). Self-modal attention, applied to speech and audio in works such as those by Pan et al. [320] and Ho et al. [321], or through cascaded multi-head designs [120], reduces intra-modal noise and sharpens temporal focus, but is fundamentally limited to improving what a single modality can express. Cross-modal attention, introduced by Krishna et al. [121] and Priyasad et al. [112] and refined through residual and cascade structures [322, 323], enables one modality to selectively attend to complementary cues in another, directly addressing the integration problem that self-modal attention cannot touch. Spatial attention identifies discriminative visual regions for image-text emotion recognition [310], while temporal attention captures emotionally salient moments in speech and video [311, 312]. Advanced architectures such as Phy-FusionNet [324] combine temporal attention with memory-augmented periodic modeling to capture long-range affective dynamics that standard attention windows miss. The key insight that emerges from surveying these mechanisms together is that they are not interchangeable solutions to the same problem but complementary instruments targeting distinct failure modes: self-modal attention addresses noise, cross-modal attention addresses integration, and spatial-temporal attention addresses localization.

Contextual attention extends this logic further by modeling discourse-level dependencies that isolated utterance attention cannot capture. Multi-EMO [98] introduces correlation-aware attention to jointly model modality interaction and dialogue flow across turns, while knowledge-aware and Bayesian co-attention frameworks [325] incorporate external or latent context to guide attention allocation toward emotionally relevant evidence. Joint Transformer-attention designs [326, 327] and MSER [328] unify temporal dependency modeling with cross-modal synchronization in a single pass. Conv-attention adapters and attention-enhanced LLM-based systems [329] further demonstrate improved adaptability under limited supervision by injecting attention-derived context into pretrained representations. The convergence of MER systems toward multi-granularity attention designs, combining self-modal, cross-modal, spatial-temporal, and contextual mechanisms within a single architecture, reflects an implicit recognition that robust emotion understanding requires all of these capabilities simultaneously. The open challenge is not adding more attention heads but learning when to rely on each granularity, a form of meta-attention that the field has not yet addressed systematically.

5.3.5 Transformer-based Architectures. Transformers [330] have become the dominant paradigm in MER, but their advantage over CNN-RNN systems goes beyond parallelism and long-range dependency modeling. The deeper architectural contribution is that cross-attention makes modality alignment and cross-modal interaction simultaneous rather than sequential: each modality can directly query representations from others at every layer, rather than

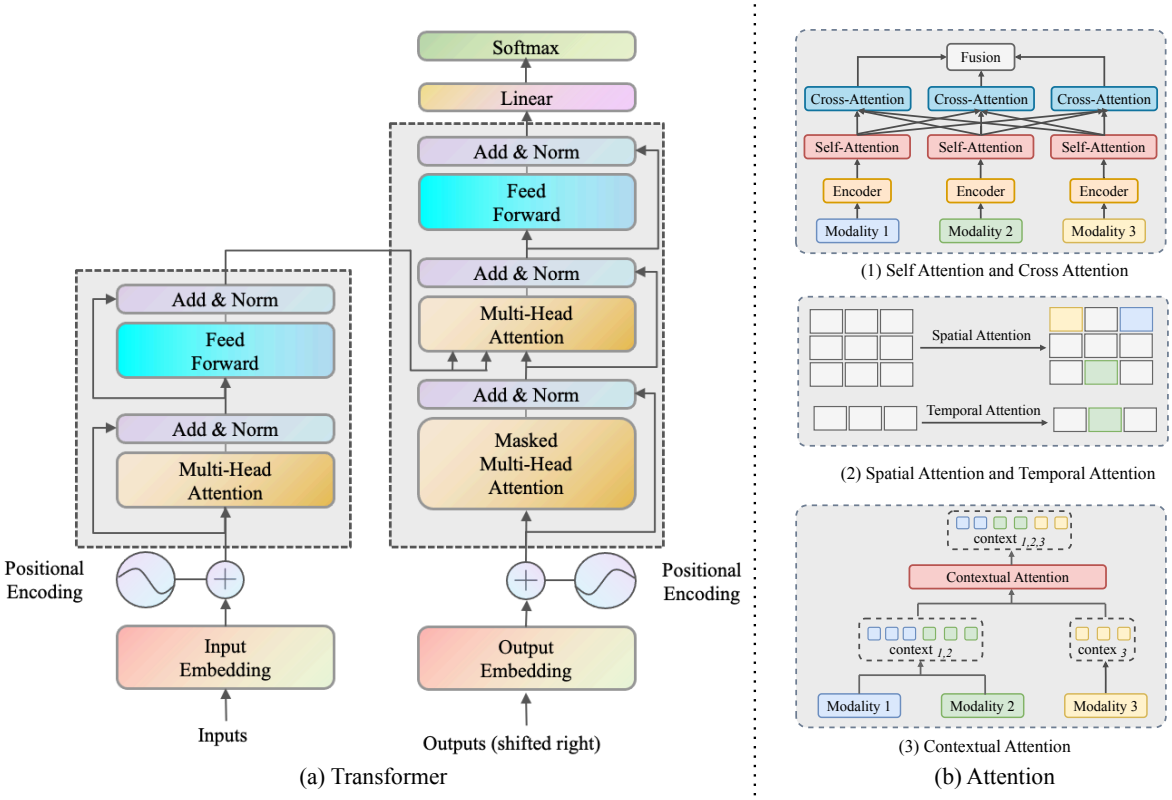


Fig. 9. (a) transformer, and (b) cross-modality attention architectures in multi-modal emotion recognition.

waiting for a downstream fusion stage. Early works demonstrated this by replacing recurrent encoders with self-supervised Transformer fusion [313, 331], and cross-modality Transformers [332–334] subsequently showed that this tighter coupling improves robustness to noise and modality inconsistency as a direct structural consequence, not merely an empirical side effect. Pre-trained audio-visual Transformers [314, 315] further improved generalization by leveraging large-scale multi-modal data, while hierarchical Transformer fusion architectures [99, 335, 336] integrated multi-level representations to capture emotional cues at different temporal scales.

A prominent subsequent direction introduced modality-aware and adaptive fusion, where models such as those proposed by Zou et al. [337] and DQ-based approaches [338] dynamically prioritize the most informative modality during inference, effectively internalizing into the architecture what earlier systems handled through post-hoc reweighting. Unified Transformer frameworks [106, 316] generalized this across multiple MER tasks within a single architecture, while flexible-input designs support continuous emotion labels, multi-label learning, and conversational settings [251, 326, 339–343]. More recent advances push toward robustness and real-world applicability: modality-collaborative Transformers with feature reconstruction [344], memory-augmented and periodicity-aware designs via Phy-FusionNet [324], capsule graph Transformers [345], adaptive cross-modal fusion networks [346], and diffusion-enhanced Transformers for conversational MER [299] each address a specific deployment gap that vanilla Transformers leave open. The critical unresolved question across this body of work is whether the performance gains of Transformer-based MER derive from the architecture itself or from the large pretrained representations it exploits: disentangling these contributions has significant implications for where future research effort should be directed.

5.3.6 Generative Architectures. Generative models occupy a distinctive role in MER because they address two problems that discriminative architectures cannot solve by design. The first is data scarcity: GAN-based augmentation [123, 347] improves generalization by synthesizing realistic audio-visual samples that diversify the training distribution beyond what real labeled data alone provides, and adversarial learning frameworks such as MALN [348] and M^{4}SER [349] promote modality-invariant, yet emotion-discriminative representations by treating cross-modal alignment as an adversarial objective. The second and more fundamental problem is missing-modality robustness: discriminative models have no principled response when a modality is absent at inference time, but generative models can reconstruct plausible representations of missing streams conditioned on available ones. Early autoencoder-style reconstruction [317, 350] established the feasibility of this approach, while diffusion-based frameworks including IMDer [115], the model by Tian et al. [351], and DiffuFuse [352] treat incomplete multi-modal inputs as a structured generative recovery problem, enabling substantially more calibrated reconstruction than masked or zero-padded alternatives.

The critical insight that separates more effective from less effective generative MER systems is that perceptual plausibility and affective coherence are not the same objective. A reconstructed speech stream may sound realistic while carrying the wrong emotional prosody, in which case it actively misleads the fusion module rather than assisting it. This requires generative objectives to be explicitly coupled with emotion-discriminative constraints, as seen in adversarial-generative frameworks [353, 354] that jointly optimize reconstruction fidelity and downstream emotion accuracy. Progressive reconstruction approaches [355] address this by incrementally conditioning generation on available affective evidence, improving coherence at each stage. The broader implication for MER system design is that generative and discriminative components should not be treated as separate modules but as co-dependent parts of a single inference pipeline, and the coupling mechanism between them remains an underexplored design principle with significant practical consequences.

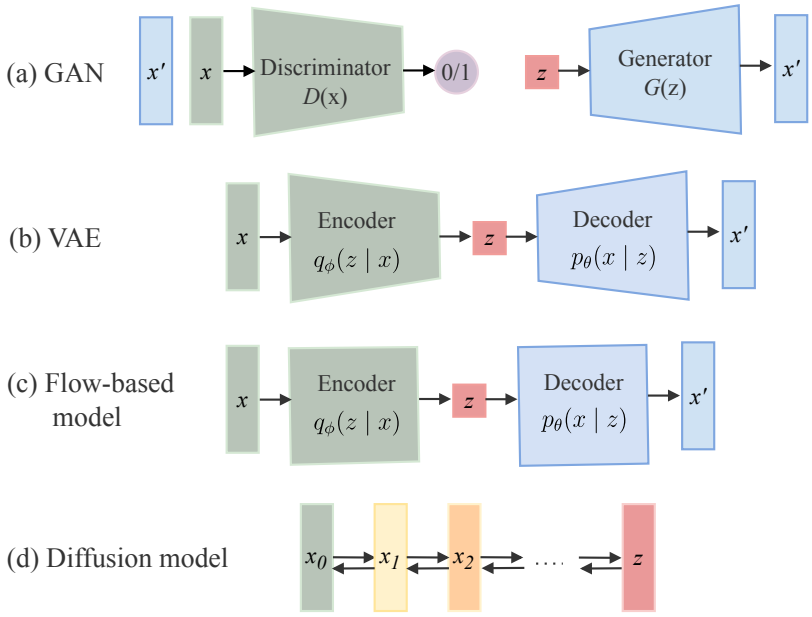


Fig. 10. (a) GAN, (b) VAE, (c) Flow-based model, and (d) Diffusion model as generative architectures in multi-modal emotion recognition.

5.3.7 Large Language Model-Based Architectures. Large language models have introduced a qualitatively different paradigm for MER by shifting the locus of multimodal understanding from task-specific fusion modules toward general-purpose reasoning over heterogeneous inputs. Unlike prior architectures that require carefully engineered cross-modal interaction mechanisms, LLM-based approaches treat emotion recognition as a language-mediated reasoning problem, where acoustic, visual, and textual cues are jointly interpreted within a unified generative framework. EmoLLM [124] and Emotion-LLaMA [318] exemplify this shift by demonstrating that instruction-tuned LLMs can not only recognize emotions across modalities but also generate natural language explanations of their predictions, a capability that purely discriminative architectures cannot provide. The critical insight is that this reasoning capacity comes without task-specific supervision: LLMs can generalize to unseen emotion categories and novel modality combinations in a way that conventional deep learning models, trained on fixed label sets and fixed fusion topologies, fundamentally cannot.

Building on this foundation, subsequent work has extended LLM-based MER in two complementary directions: conversational understanding and open-vocabulary generalization. DialogueMLLM [356] addresses the challenge of recognizing emotions in dynamic conversational contexts by instruction-tuning LLMs on dialogue-level dependencies, capturing how emotional states shift across turns in ways that utterance-level models miss. Additionally, Zhang et al. [357] introduced modality sabotage, a lightweight failure mode in LLMs where a high-confidence unimodal error dominates other evidence and misleads multimodal fusion. AffectGPT [133] pushes further by introducing explainability as a first-class objective, providing reasoning traces that make emotional predictions interpretable in real-world applications. On the generalization front, OV-MER [128] and GPT-4V-based approaches [248] demonstrate meaningful zero-shot performance on emotion categories unseen during training, suggesting that the semantic richness of LLM pretraining encodes affective knowledge that transfers across domains without retraining. AffectGPT-R1 [133] and R1-Omni [134] further integrate reinforcement learning to improve both open-vocabulary performance and interpretability, combining the reasoning strengths of LLMs with reward-guided optimization. Most recently, OMNISAPIENS-7B [358] introduces a unified benchmark for multimodal human behaviour understanding, and OMNISAPIENS-7B 2.0 [359] further advances this direction with reinforcement learning to balance learning across heterogeneous modalities and tasks.

Despite these advances, LLM-based MER still faces several challenges. First, existing methods remain sensitive to prompt variations, such as changes in separator tokens, wording, and the ordering of emotion labels [360]. This limitation suggests that LLMs may not fully comprehend human emotions. Furthermore, overemphasizing LLMs' emotion understanding capabilities could compromise their general-purpose performance [361]. How to effectively mitigate memory forgetting also requires further investigation. In addition, multimodal hallucination remains a significant problem in LLMs and negatively affects the reliability of their emotion prediction results [362]. Consequently, considerable work remains to be done to advance LLM-based MER.

5.4 Summary of Multi-modal Emotion Recognition

Multi-modal emotion recognition has evolved from modality concatenation and kernel-based fusion toward deeply integrated, context-aware, and reasoning-capable architectures. Reviewing the seven dimensions examined in this section reveals a unifying pattern: the field has progressively moved from treating fusion as a feature engineering problem to treating it as a structured inference problem, where modality alignment, dominance, complementarity, and robustness are not separate challenges to be patched independently but interdependent properties of a single joint representation. The rise of Transformer and LLM-based architectures reflects this shift most visibly, but the enduring open challenges of benchmark-to-deployment generalization, demographic fairness, prompt sensitivity, and the annotation bottleneck are architecture-agnostic and will require advances in data, evaluation protocol, and learning paradigm alongside continued architectural innovation.

6 FAIRNESS AND ETHICAL IMPLICATIONS

Automatic emotion recognition in conversations offers powerful tools for applications such as healthcare, education, and human–computer interaction. However, these systems risk amplifying existing biases if fairness and ethical considerations are neglected. Previous work are strongly dependent on

targeted datasets skewed toward specific demographics or culture groups, which may lead to uneven performance across gender, ethnicity, or age[363]. Many studies have shown that emotion perception itself varies with culture and social context, making it critical for automatic emotion recognition in conversations to incorporate fairness-aware learning objectives and diverse, representative datasets. For example, the same tone or facial expression can convey different emotions across cultural groups, and conversational norms such as silence, politeness, or sarcasm are also culturally shaped. These variations highlight that models trained on narrow populations may systematically misinterpret emotional cues, reinforcing biases, or excluding marginalized voices (see Fig. 11).

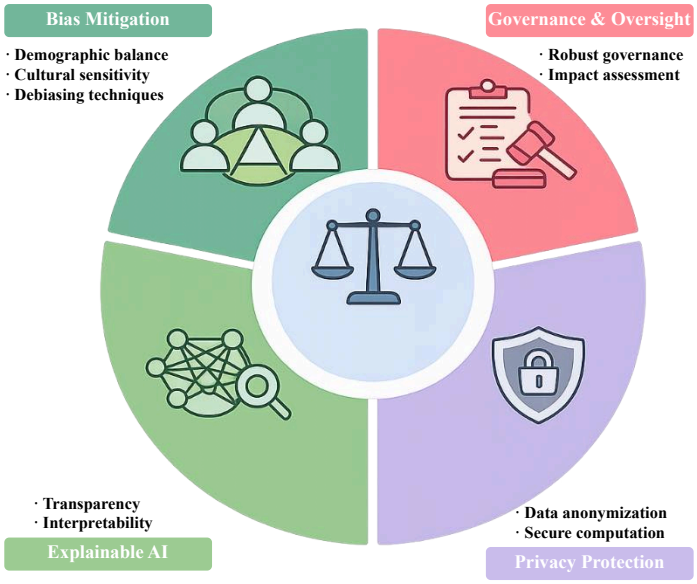


Fig. 11. Fairness and ethical implications in emotion recognition.

Beyond data imbalance, interpretability and transparency also raise ethical concerns. Users should be able to understand how conversational cues are processed, and how predictions are derived. Although contextualized reasoning networks [118, 125] and knowledge-aware graph approaches [131] show promise, their internal decision-making remains opaque. Explainable emotion recognition techniques, such as attention visualization or concept-based explanations, are essential for accountability. Moreover, without interpretable reasoning, it becomes difficult to diagnose systematic biases or erroneous predictions, especially in sensitive contexts. Therefore, explainability is not merely an auxiliary feature but a critical requirement for establishing trust and enabling responsible deployment. Recent work increasingly emphasizes human-centered interpretability, where explanations are tailored to different stakeholders, including developers, domain experts, and end-users. Such approaches highlight that transparency should not only describe how a model works but also support meaningful oversight and informed decision-making.

Meanwhile, privacy risks emerge when conversational data is collected without explicit consent, highlighting the importance of robust governance protocols and privacy-preserving computation. Importantly, fairness and ethical issues surrounding emotion analysis are not limited to conversational settings. Similar concerns have also been reported in related affective technologies, including facial expression recognition, physiological sensing, and affect detection used in surveillance, recruitment, and educational analytics [68]. Previous models may disproportionately misclassify expressions exhibited by certain racial or cultural groups, or systematically interpret atypical behavioral patterns as negative affect. When such biased models are applied in consequential domains such as hiring decisions, student assessment, or public security, they may unintentionally legitimize intrusive monitoring and reinforce structural inequalities. In these contexts, incorrect emotional judgments can influence access to opportunities, shape institutional treatment of individuals, and affect how society defines and values different emotional expressions. These observations underscore that ethical challenges extend beyond model training to issues of governance, social impact, and human autonomy.

In high-stakes environments such as education, healthcare, public security, or workplace monitoring, misclassification of emotions can have serious consequences for well-being, trust, legal accountability, and fairness. Ethical frameworks thus call for ongoing bias audits, participatory design with end-users, and culturally adaptive interpretations of emotional cues. Techniques such as federated learning, counterfactual fairness constraints, debiasing objectives, and culturally calibrated benchmarks may help mitigate risks. Moreover, insights from other domains, including fairness principles applied in recommendation systems, predictive policing, and hiring algorithms, can meaningfully inform the ethical design of conversational emotion recognition. Beyond methodological advances, interdisciplinary collaborations with social scientists, ethicists, and affected communities are becoming increasingly important for grounding technical solutions in lived experiences. Continued evaluation of system impact in real-world settings is also essential, as fairness cannot be assumed at deployment but requires sustained monitoring and refinement over time.

In summary, fairness and ethical considerations are integral to the responsible advancement of emotion recognition research. The field is evolving from a focus on limited evaluation criteria toward more inclusive, transparent, and privacy-preserving systems designed to reflect demographic, cultural, and contextual diversity. Future progress depends on integrating fairness-aware objectives, explainable reasoning mechanisms, cross-domain ethical frameworks, and governance models that ensure emotion recognition technologies enhance human well-being rather than reinforce inequity.

7 FUTURE DIRECTIONS

The preceding sections have surfaced a set of concrete, recurring limitations across emotion modeling, dataset construction, uni-modal learning, multimodal fusion, and evaluation practice. Rather than restating generic algorithmic aspirations, the directions below are each anchored to a specific gap

identified in our analysis, with the aim of translating the unified framework proposed in this survey into actionable research priorities. Fig. 12 situates these directions within the broader vision of affective foundation models.

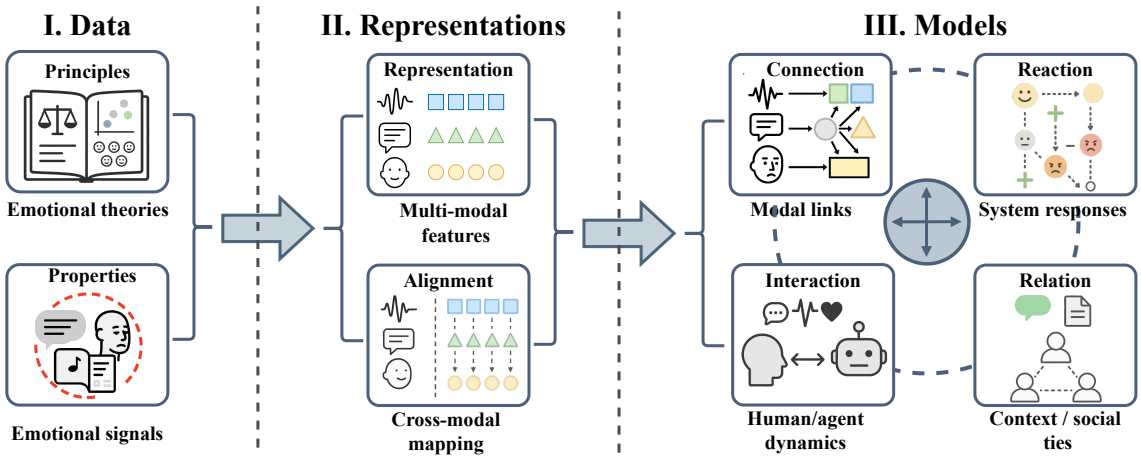


Fig. 12. Exploring data, representations, and models helps establish a strong theoretical and practical foundation for the affective model.

Appraisal-Based Modeling as an Underexplored Bridge for Conversational Emotion: As established in Section 2, the dominant emotion frameworks used in practice remain Ekman's categorical model and the valence–arousal dimensional space, despite their well-documented limitations in capturing the dynamic, relational character of emotions in conversation. Appraisal-based models—which explain emotion as the outcome of a speaker's evaluation of novelty, goal relevance, and coping potential [51, 52]—offer a theoretically richer account of *why* emotions shift across turns. Yet, as our review of conversational MER in Section 5 reveals, no existing dataset or benchmark has operationalized appraisal dimensions at the turn level. Future work should therefore (i) annotate existing conversational corpora such as IEMOCAP [62] and MELD [63] with appraisal-derived labels—even at a coarse three-level granularity—and (ii) design model architectures that condition emotion prediction on inferred speaker goals and situational context, rather than treating each utterance as an independent classification instance. This would transform conversational MER from a sequence-labeling task into a richer form of affective reasoning grounded in cognitive theory.

Establishing When and Why Fusion Strategies Succeed or Fail: Section 5 and Table 5 catalog a wide range of fusion architectures—early, late, model-level, and hybrid—but the field currently lacks principled criteria for selecting among them. Our cross-study analysis reveals that performance differences between fusion strategies are highly sensitive to modality availability, dataset size, and label granularity, yet these factors are rarely controlled for in published comparisons. Future benchmarks should systematically test models under varying modality-availability conditions—including unimodal-only baselines, full trimodal inputs, and missing-modality scenarios—so that the community can develop empirically grounded rules for fusion strategy selection. In particular, future work should investigate the conditions under which late fusion outperforms cross-modal attention despite its architectural simplicity, and should quantify the marginal gain of adding physiological signals (P) to audio–text–visual (ATV) systems, a question that Table 1 shows remains systematically unanswered in the literature.

Targeted Robustness Benchmarking Aligned with Table 4 Generalization Gaps: Our synthesis of uni-modal and multi-modal models in Sections 4–5 reveals a consistent pattern: state-of-the-art results are reported on a small cluster of benchmarks (IEMOCAP, MELD, CMU-MOSI), and cross-corpus evaluations are the exception rather than the rule. The models surveyed in Table 4 rarely report performance on held-out domains, languages, or demographic subgroups, making it impossible to distinguish genuine generalization from dataset-specific overfitting. Future evaluation protocols should mandate at minimum: (i) leave-one-corpus-out evaluation across at least two datasets sharing a compatible label space; (ii) disaggregated performance reporting by speaker gender, age group, and language; and (iii) robustness scores under controlled noise injection into each modality. Establishing these as community norms—analogous to out-of-distribution benchmarks in computer vision—would make the generalization claims implicit in Table 4 empirically testable rather than assumed.

Scalable and Privacy-Preserving Data Strategies to Address Dataset Fragmentation: Section 3 documents that existing emotion corpora are fragmented across incompatible label spaces, skewed toward acted or induced elicitation paradigms, and severely underrepresented in non-English and non-Western cultural contexts. This fragmentation is not merely inconvenient—it is a direct cause of the cross-corpus generalization failures identified above. Future data strategies should pursue three complementary approaches: (i) semi-supervised and self-supervised annotation pipelines [138] that can scale naturalistic data collection without full manual labeling [364]; (ii) cross-cultural benchmark consortia analogous to SEWA [50] but covering a broader set of languages and interaction styles; and (iii) federated data collection frameworks [130] that allow institutions to contribute data while preserving participant privacy—particularly important for clinical and educational deployment contexts.

Interpretable and Domain-Generalizable Representation Learning: As highlighted in Sections 4–5, the shift from hand-crafted features to large pre-trained models (wav2vec 2.0, HuBERT, WavLM, ViT) has substantially improved within-dataset accuracy, but has also reduced model interpretability and increased sensitivity to domain shift. The models reviewed exhibit significant performance degradation when applied outside their training distribution, yet few incorporate explicit domain-adaptation or continual-learning mechanisms [4, 28]. Future research should develop hybrid architectures that combine the representational power of self-supervised pre-training with modular, interpretable components—such as attention over acoustic landmarks or facial action unit activations—that allow practitioners to audit model behavior and identify failure modes without full retraining [18, 365].

Fairness-Aware Evaluation as a First-Class Benchmark Criterion: Section 6 identifies that demographic and cultural bias in both training data and evaluation protocols remains a largely unresolved issue. Crucially, our survey of existing benchmarks reveals that fairness metrics are almost never reported alongside accuracy metrics, meaning that a model achieving high weighted-F1 on IEMOCAP may perform far worse for specific speaker demographics without this being detected. Future benchmarks should treat equitable performance across demographic subgroups as a first-class evaluation

criterion, on par with overall accuracy. Concretely, this requires datasets annotated with speaker demographic metadata, evaluation scripts that report performance gaps as standard outputs, and challenge tracks—analogous to those in MER2024 [66] and MER2025 [67]—that explicitly reward fairness alongside accuracy. Since emotional expression varies systematically across cultural contexts [129], cross-cultural calibration of annotation schemes should also become standard practice, and holistic affective computing frameworks that jointly model emotions alongside related states such as stress and cognitive load should be explored [142].

Toward Affective Foundation Models with Rigorous Generalization Criteria: Recent advances in large language and vision–language models have catalyzed interest in unified affective models capable of few-shot adaptation and open-vocabulary emotion reasoning [128, 133, 135]. However, the evaluation of these models currently lacks the rigor applied to task-specific systems: reported results mix zero-shot, few-shot, and fine-tuned settings without standardized protocols, and generalization to non-English or non-acted data is rarely tested. Future work on affective foundation models should therefore be accompanied by a dedicated evaluation suite—analogous to EmoBench [141]—that assesses: (i) zero-shot transfer across label spaces; (ii) compositional reasoning over multi-turn affective context; (iii) calibration under distribution shift; and (iv) alignment with human annotations from diverse cultural and linguistic backgrounds. Only with such criteria can the field determine whether affective foundation models represent genuine scientific progress or primarily reflect the statistical biases of their pre-training corpora.

In summary, the future of emotion recognition is not simply a matter of scaling models or collecting more data. The specific gaps identified in this survey—appraisal-theoretic modeling of conversational dynamics, principled fusion strategy selection, systematic cross-corpus generalization testing, culturally inclusive data strategies, interpretable robustness under domain shift, fairness-aware benchmarking, and rigorous evaluation of affective foundation models—define a concrete research agenda. Progress on these fronts will require the community to prioritize reproducibility and comparability over benchmark-specific performance, and to treat theoretical grounding, ethical accountability, and cross-disciplinary collaboration [219, 366, 367] not as supplementary concerns but as integral components of methodological rigor.

8 CONCLUSION

This survey has presented a unified synthesis of deep learning-based emotion recognition, jointly examining unimodal and multimodal approaches within a coherent analytical framework spanning emotion modeling, dataset curation, modality-specific representation learning, fusion strategy design, and evaluation. Across all three unimodal modalities, progress has been substantial: facial expression recognition has evolved from hand-crafted geometric descriptors toward spatio-temporal and Transformer-based architectures; speech emotion recognition has been transformed by self-supervised pretraining, which resolves the fundamental mismatch between model capacity and the scarcity of emotion-labeled data; and textual emotion recognition has advanced through pretrained language models, though the central challenge of capturing how emotional meaning evolves across conversational turns remains open. At the multi-modal level, fusion research has progressed from simple feature concatenation toward principled strategies that treat modality alignment, dominance, complementarity, and robustness as explicit modeling objectives, and large language model-based architectures have introduced a qualitatively different paradigm by enabling zero-shot generalization and natural language explainability. Yet several foundational challenges persist across the field: label ambiguity and annotation subjectivity undermine cross-dataset comparability, evaluation protocols remain fragmented and poorly equipped to detect models that are functionally uni-modal despite nominal multi-modal fusion, and demographic fairness across speaker gender, age, and dialect is consistently acknowledged but rarely addressed in a principled way. These limitations are structural rather than incidental, and future progress will require not only stronger models but also more ecologically valid datasets, standardized evaluation frameworks, and systematic treatment of the ethical implications of affective inference in real-world deployment.

REFERENCES

[1] Yan Wang, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang, and Zhongxue Gan. A survey on facial expression recognition of static and dynamic emotions. *arXiv preprint arXiv:2408.15777*, 2024.

[2] Mohan Karnati, Ayan Seal, Debotosh Bhattacharjee, Anis Yazidi, and Ondrej Krejcar. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Transactions on Instrumentation and Measurement*, 72:1–31, 2023.

[3] Guimin Hu, Yi Xin, Weimin Lyu, Haojian Huang, Chang Sun, Zhihong Zhu, Lin Gui, Ruichu Cai, Erik Cambria, and Hasti Seifi. Recent trends of multimodal affective computing: A survey from nlp perspective. *arXiv preprint arXiv:2409.07388*, 2024.

[4] Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *arXiv preprint arXiv:2505.20511*, 2025.

[5] Xiaowei Zhang, Jinyong Liu, Jian Shen, Shaojie Li, Kechen Hou, Bin Hu, Jin Gao, and Tong Zhang. Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. *IEEE transactions on cybernetics*, 51(9):4386–4399, 2020.

[6] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction. *Frontiers in Computer Science*, 3:767767, 2021.

[7] Guimin Hu, Yi Zhao, and Guangming Lu. Improving representation with hierarchical contrastive learning for emotion-cause pair extraction. *IEEE Transactions on Affective Computing*, 15(4):1997–2011, 2024.

[8] Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. UniMEEC: Towards unified multimodal emotion recognition and emotion cause. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5248–5261, 2024.

[9] AV Geetha, T Mala, D Priyanka, and E Uma. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024.

[10] Sepideh Kalateh, Luis A Estrada-Jimenez, Sanaz Nikghadam-Hojjati, and Jose Barata. A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. *IEEE Access*, 12:103976–104019, 2024.

[11] Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information fusion*, 102:102019, 2024.

[12] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237:121692, 2024.

[13] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[14] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing*, 14(1):68–88, 2020.

[15] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95: 306–325, 2023.

[16] Samuel Kakuba, Alwin Poulose, and Dong Seog Han. Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model. *IEEE Access*, 11:113769–113789, 2023.

[17] Kaouther Ezzameli and Hela Mahersia. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847, 2023.

[18] Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440, 2023.

[19] Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502*, 2025.

[20] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10):1163, 2021.

[21] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73, 2021.

[22] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[23] Youddha Beer Singh and Shivani Goel. A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492:245–263, 2022.

[24] Ruhina Karani and Sharmishta Desai. Review on multimodal fusion techniques for human emotion recognition. *Int. J. Adv. Comput. Sci. Appl*, 13:287–296, 2022.

[25] M Maithri, U Raghavendra, Anjan Gudigar, Jyothi Samanth, Prabal Datta Barua, Murugappan Murugappan, Yashas Chakole, and U Rajendra Acharya. Automated emotion recognition: Current trends and future perspectives. *Computer methods and programs in biomedicine*, 215:106646, 2022.

[26] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

[27] Guoying Zhao, Xiaobai Li, Yante Li, and Matti Pietikäinen. Facial micro-expressions: An overview. *Proceedings of the IEEE*, 111(10):1215–1235, 2023.

[28] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67, 2021.

[29] Mohammed Jawad Al-Dujaili and Abbas Ebrahimi-Moghadam. Speech emotion recognition: a comprehensive survey. *Wireless Personal Communications*, 129(4): 2525–2561, 2023.

[30] Javier de Lope and Manuel Graña. An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11, 2023.

[31] Ahlam Hashem, Muhammad Arif, and Manal Alghamdi. Speech emotion recognition approaches: A systematic review. *Speech Communication*, page 102974, 2023.

[32] Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L Schneider. Speech emotion recognition using machine learning—a systematic review. *Intelligent systems with applications*, 20:200266, 2023.

[33] Bagus Tris Atmaja, Akira Sasou, and Masato Akagi. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*, 140:11–28, 2022.

[34] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023.

[35] Yujian Cai, Xingguang Li, and Jinsong Li. Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. *Sensors*, 23 (5):2455, 2023.

[36] Bei Pan, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561:126866, 2023.

[37] T Kopalidis, V Solachidis, N Vretos, and P Daras. Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. information, 15 (3), 135, 2024.

[38] Tarun Rathi and Manoj Tripathy. Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review. *Speech Communication*, 162:103102, 2024.

[39] Priyanka Thakur, Nirmal Kaur, Naveen Aggarwal, and Sarbjeet Singh. A comprehensive review of unimodal and multimodal emotion detection: Datasets, approaches, and limitations. *Expert Systems*, 42(9):e70103, 2025.

[40] Ghulam Muhammad, Sumayah Almuntasheri, Fadia Alenezi, Nooran Alhadi, and Victor CM Leung. Eeg-based multimodal emotion recognition: Recent progress, challenges, and future directions. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[41] MJ Dileep Kumar, M Sukesh Rao, and KC Narendra. Multimodal emotion recognition: A comprehensive survey of datasets, methods, and applications. *IEEE Access*, 13:201067–201097, 2025.

[42] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwook Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics, 2020.

[43] Adil Chakhtouna, Sara Sekkate, and Abdellah Adib. Speech emotion recognition: A systematic mega-review of techniques and pipelines. *Information Fusion*, page 104161, 2026.

[44] Mohamed Abdeldayem, Hesham FA Hamed, and Amr M Nagy. Facial expression recognition: A survey of techniques, datasets, and real-world challenges. *Statistics, Optimization & Information Computing*, 15(1):733–761, 2026.

[45] Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotionthinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. *ICLR*, 2026.

[46] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[47] Robert Plutchik. *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association, 2003.

[48] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[49] Albert Mehrabian and James A Russell. *An approach to environmental psychology*. the MIT Press, 1974.

[50] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1022–1040, 2019.

[51] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[52] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

[53] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[54] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[55] Kate Dupuis and M Kathleen Pichora-Fuller. Toronto emotional speech set (tess)-younger talker_happy. 2010.

[56] Felix Burkhardt, Oliver Schrüfer, Uwe Reichel, Hagen Wierstorf, Anna Derington, Florian Eyben, and Björn Schuller. Emodb 2.0: A database of emotional speech in a world that is not black or white but grey. In *Proc. of Interspeech*, pages 4488–4492, 2025.

[57] Klaus R Scherer and Harald G Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310, 1994.

[58] Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, 2017.

[59] Steven R Livingstone and Frank A Russo. 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*, 13(5):e0196391, 2018.

[60] Sanaul Haq. Speaker-dependent audio-visual emotion recognition. *personal. ee. surrey. ac. uk*, 2009.

[61] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd international conference on data engineering workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[62] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[63] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536, 2019.

[64] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[65] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 9610–9614, 2023.

[66] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, 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*, pages 41–48, 2024.

[67] Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13837–13842, 2025.

[68] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone D Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):233, 2020.

[69] K Slimani, Mohamed Kas, Youssef El Merabet, Rochdi Messoussi, and Yassine Ruichek. Facial emotion recognition: A comparative analysis using 22 lbp variants. In *Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pages 88–94, 2018.

[70] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[71] Tong Liu, Jing Li, Jia Wu, Bo Du, Jun Chang, and Yi Liu. Facial expression recognition on the high aggregation subgraphs. *IEEE Transactions on Image Processing*, 32: 3732–3745, 2023.

[72] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 101–110, 2022.

[73] Weicong Chen, Dong Zhang, Ming Li, and Dah-Jye Lee. Stcam: Spatial-temporal and channel attention module for dynamic facial expression recognition. *IEEE Transactions on Affective Computing*, 14(1):800–810, 2020.

[74] Bowen Shi, Yue Zhang, Jie Yang, and Yao Zhao. Stam: Spatiotemporal transformer with attention modulation for video-based emotion recognition. In *ACM MM*, 2022.

[75] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023.

[76] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny Lo. Revisiting self-supervised contrastive learning for facial expression recognition. *arXiv preprint arXiv:2210.03853*, 2022.

[77] Yuanyuan Liu, Wenbin Wang, Yibing Zhan, Shaoze Feng, Kejun Liu, and Zhe Chen. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9717–9728, 2023.

[78] Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. Svfap: Self-supervised video facial affect perceiver. *IEEE Transactions on Affective Computing*, 16(1):405–422, 2024.

[79] Licai Sun, Xingxun Jiang, Haoyu Chen, Yante Li, Zheng Lian, Biu Liu, Yuan Zong, Wenming Zheng, Jukka M Leppänen, and Guoying Zhao. Learning transferable facial emotion representations from large-scale semantically rich captions. *arXiv preprint arXiv:2507.21015*, 2025.

[80] GH Mohmad Dar and Radhakrishnan Delhibabu. Speech databases, speech features, and classifiers in speech emotion recognition: A review. *IEEE Access*, 12:151122–151152, 2024.

[81] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[82] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[83] Shuaiqi Chen, Xiaofen Xing, Weibin Zhang, Weidong Chen, and Xiangmin Xu. Dwformer: Dynamic window transformer for speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[84] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[85] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[86] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[87] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, 2024.

[88] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):1711–1724, 2024.

[89] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[90] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *National Research Council, Canada*, 2:234, 2013.

[91] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[92] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[93] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, June 2018.

[94] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[95] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[96] Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*, 2023.

[97] Yunhe Xie, Cheng-Jie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. A dual contrastive learning framework for enhanced multimodal conversational emotion recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4055–4065, 2025.

[98] Tao Shi and Shao-Lun Huang. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, 2023.

[99] Zhengdao Zhao, Yuhua Wang, Guang Shen, Yuezhu Xu, and Jiayuan Zhang. Tdfnet: Transformer-based deep-scale fusion network for multimodal emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3771–3782, 2023.

[100] Jiahui Pan, Weijie Fang, Zhihang Zhang, Bingzhi Chen, Zheng Zhang, and Shuihua Wang. Multimodal emotion recognition based on facial expressions, speech, and eeg. *IEEE Open Journal of Engineering in Medicine and Biology*, 5:396–403, 2023.

[101] Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19089–19097, 2024.

[102] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[103] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, 2018.

[104] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020.

[105] Jiachen Luo, Huy Phan, and Joshua Reiss. Cross-modal fusion techniques for utterance-level emotion recognition from text and speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[106] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, 2022.

[107] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1103–1114, 2017.

[108] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang Yamada, Amir Zadeh, and Louis-Philippe Morency. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569. Association for Computational Linguistics, 2019.

[109] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2023.

[110] ChenYuan He, Senbin Zhu, Hongde Liu, Fei Gao, Yuxiang Jia, Hongying Zan, and Min Peng. Dialoguemmt: Dialogue scenes understanding enhanced multi-modal multi-task tuning for emotion recognition in conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2497–2512. Association for Computational Linguistics, 2025.

[111] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2359–2369, 2020.

[112] Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Attention driven fusion for multi-modal emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3227–3231. IEEE, 2020.

[113] Ming Xu, Tuo Shi, Hao Zhang, Zeyi Liu, and Xiao He. A hierarchical cross-modal spatial fusion network for multimodal emotion recognition. *IEEE Transactions on Artificial Intelligence*, 2025.

[114] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021.

[115] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[116] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

[117] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524, 2013.

[118] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.

[119] Yazhou Zhang, Ao Jia, Bo Wang, Peng Zhang, Dongming Zhao, Pu Li, Yuexian Hou, Xiaojia Jin, Dawei Song, and Jing Qin. M3gat: A multi-modal, multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Transactions on Information Systems*, 42(1):1–32, 2023.

[120] Jiahao Zheng, Sen Zhang, Zilu Wang, Xiaoping Wang, and Zhigang Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE transactions on multimedia*, 25:2213–2225, 2022.

[121] Dushyant N Krishna and Ankita Patil. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *Interspeech*, pages 4243–4247, 2020.

[122] Zheng Lian, Bin Liu, and Jianhua Tao. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000, 2021.

[123] Yun Luo, Li-Zhen Zhu, and Bao-Liang Lu. A gan-based data augmentation method for multimodal emotion recognition. In *International Symposium on Neural Networks*, pages 141–150. Springer, 2019.

[124] Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*, 2024.

[125] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4148–4164, 2022.

[126] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661, 2022.

[127] Geng Tu, Feng Xiong, Bin Liang, and Ruifeng Xu. 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*, pages 2266–2270, 2024.

[128] Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. In *International Conference on Machine Learning*, pages 37015–37050. PMLR, 2025.

[129] James A Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.

[130] Neha Gahlan and Divyashikha Sethia. Federated learning in emotion recognition systems based on physiological signals for privacy preservation: a review. *Multimedia Tools and Applications*, 84(13):12417–12485, 2025.

[131] Lei Zhang et al. Context- and knowledge-aware graph convolutional network for multimodal emotion recognition. In *Proceedings of AAAI*, pages 1–9, 2022.

[132] Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, et al. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023.

[133] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *International Conference on Machine Learning*, pages 36993–37014. PMLR, 2025.

[134] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

[135] Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, and Xiaobai Li. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025.

[136] Zheng Lian, Licai Sun, Lan Chen, Haoyu Chen, Zebang Cheng, Fan Zhang, Ziyu Jia, Ziyang Ma, Fei Ma, Xiaojiang Peng, et al. Emoprefer: Can large language models understand human emotion preferences? *ICLR*, 2026.

[137] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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 (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[138] Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2024.

[139] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

[140] Lucas Goncalves and Carlos Busso. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170, 2022.

[141] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, 2024.

[142] Yuanchao Li, Dimitrios Kollias, Guillaume Chanel, Marios Fanourakis, Michal Muszynski, Brandon M. Booth, Leimin Tian, Madhawa Perera, Catherine Lai, and Huili Chen. Hrai 2025: The 1st workshop on holistic and responsible affective intelligence. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pages 814–817, 2025.

[143] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[144] James A Russell. Measures of emotion. In *The measurement of emotions*, pages 83–111. Elsevier, 1989.

[145] Gerben A Van Kleef and Stéphane Côté. The social effects of emotions. *Annual review of psychology*, 73:629–658, 2022.

[146] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, 1993.

[147] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[148] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[149] Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. Challenges and opportunities of text-based emotion detection: A survey. *IEEE access*, 12:18416–18450, 2024.

[150] Eva G Krumhuber, Lina I Skora, Harold CH Hill, and Karen Lander. The role of facial movements in emotion recognition. *Nature Reviews Psychology*, 2(5):283–296, 2023.
[151] Carroll E Izard. *Human emotions*. Springer Science & Business Media, 2013.
[152] John L Andreassi. *Psychophysiology: Human behavior and physiological response*. Psychology press, 2010.
[153] Matthew L Dixon, Ravi Thiruchselvam, Rebecca Todd, and Kalina Christoff. Emotion and the prefrontal cortex: An integrative review. *Psychological bulletin*, 143(10): 1033, 2017.
[154] Ling Wang, Jiayu Hao, and Tie Hua Zhou. ECG multi-emotion recognition based on heart rate variability signal features mining. *Sensors*, 23(20):8636, 2023.
[155] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343: 35–55, 2019.
[156] Wanhui Wen, Guangyuan Liu, Nanpu Cheng, Jie Wei, Pengchao Shangguan, and Wenjin Huang. Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Transactions on Affective Computing*, 5(2):126–140, 2014.
[157] Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. The japanese female facial expression (jaffe) dataset. *(No Title)*, 1998.
[158] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9):607–619, 2011.
[159] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3): 34–41, 2012.
[160] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.
[161] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016.
[162] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*, pages 137–144, 2016.
[163] Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, et al. The msp-podcast corpus. *arXiv preprint arXiv:2509.09791*, 2025.
[164] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.
[165] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.
[166] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.
[167] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.
[168] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pages 24–32, 2022.
[169] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. CHEAVD: a Chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924, 2017.
[170] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE transactions on affective computing*, 12(2):479–493, 2018.
[171] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.
[172] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, 2020.
[173] Olga Perepelkina, Evdokia Kazimirova, and Maria Konstantinova. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. In *International Conference on Speech and Computer*, pages 501–510. Springer, 2018.
[174] A Aruna Gladys and V Vetriselvi. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, 556:126693, 2023.
[175] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.
[176] Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.
[177] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.
[178] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016.
[179] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics, 2019.
[180] Xiaoyi Feng, Matti Pietikainen, and Abdenour Hadid. Facial expression recognition based on local binary patterns. *Pattern Recognition and Image Analysis*, 17(4): 592–598, 2007. doi: 10.1134/S1054661807040190.
[181] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998. doi: 10.1109/AFGR.1998.670949.
[182] Haifeng Zhang, Wen Su, Jun Yu, and Zengfu Wang. Identity–expression dual branch network for facial expression recognition. *IEEE transactions on cognitive and developmental systems*, 13(4):898–911, 2020.
[183] Yan Yan, Zizhao Zhang, Si Chen, and Hanzi Wang. Low-resolution facial expression recognition: A filter learning perspective. *Signal Processing*, 169:107370, 2020.
[184] Zhengning Wang, Fanwei Zeng, Shuaicheng Liu, and Bing Zeng. OAENet: Oriented attention ensemble for accurate facial expression recognition. *Pattern Recognition*, 112:107694, 2021.
[185] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.
[186] Xing Jin, Zhihui Lai, and Zhong Jin. Learning dynamic relationships for facial expression recognition based on graph convolutional network. *IEEE Transactions on Image Processing*, 30:7143–7155, 2021.
[187] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. doi: 10.1109/CVPR.2014.226.
[188] Zhenbo Yu, Guangcan Liu, Qingshan Liu, and Jiankang Deng. Spatio-temporal convolutional features with nested lstm for facial expression recognition. *Neurocomputing*, 317:50–57, 2018. doi: 10.1016/j.neucom.2018.07.028.
[189] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. SAANet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020.
[190] Rajesh Singh, Sumeet Saurav, Tarun Kumar, Ravi Saini, Anil Vohra, and Sanjay Singh. Facial expression recognition in videos using hybrid cnn & convlstm. *International Journal of Information Technology*, 15(4):1819–1830, 2023. doi: 10.1007/s41870-023-01183-0.
[191] Xing Jin, Xiyin Wu, Libo Weng, and Qiaolin Ye. Lightweight binary convolutional-transformers fusion network for facial expression recognition. *Engineering Applications of Artificial Intelligence*, 158:111315, 2025.
[192] Yunseong Cho, Chanwoo Kim, Hoseong Cho, Yunhoe Ku, Eunseo Kim, Muhammadjon Boboev, Joonseok Lee, and Seungryul Baek. Rmfer: Semi-supervised contrastive learning for facial expression recognition with reaction mashup video. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5901–5910, 2024. doi: 10.1109/WACV57701.2024.00581.
[193] Xiaoqing Wang, Xiangjun Wang, and Yubo Ni. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience*, 2018(1):7208794, 2018.
[194] Shuvendu Roy and Ali Etemad. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 253–257, 2021.

[195] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016.

[196] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[197] Zhengyin Du, Suowei Wu, Di Huang, Weixin Li, and Yunhong Wang. Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. *IEEE Transactions on Affective Computing*, 12(3):565–578, 2019.

[198] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[199] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. Vitfer: Facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, 2022. doi: 10.3390/asi5040080.

[200] Mahdi Pourmirzaei, Gholam Ali Montazer, and Farzaneh Esmaili. Using self-supervised auxiliary tasks to improve fine-grained facial representation. *arXiv preprint arXiv:2105.06421*, 2021.

[201] Rabie Helaly, Seifeddine Messaoud, Soulef Bouaafia, Mohamed Ali Hajjaji, and Abdellatif Mtibaa. DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18. *Signal, Image and Video Processing*, 17(6):2731–2744, 2023.

[202] Jun Zhang, Fei Wu, and Xin Yu. D2SP: Dual-Stage Purification for Enhanced Facial Emotion Recognition in Complex Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1089–1097, 2025.

[203] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*, 2019.

[204] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.

[205] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020*, pages 6419–6423, 2020.

[206] Adrian Bogdan Stânea, Vlad Strilețchi, Cosmin Strilețchi, and Adriana Stan. An analysis of large speech models-based representations for speech emotion recognition. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 100–104. IEEE, 2023.

[207] Haibin Wu, Huang-Cheng Chou, Kai-Wei Chang, Lucas Goncalves, Jiawei Du, Jyh-Shing Roger Jang, Chi-Chun Lee, and Hung-Yi Lee. Emo-superb: An in-depth look at speech emotion recognition. *arXiv preprint arXiv:2402.13018*, 2024.

[208] Ulku Bayraktar, Hasan Kilimci, H Hakan Kilinc, and Zeynep Hilal Kilimci. Assessing audio-based transformer models for speech emotion recognition. In *2023 7th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, pages 1–7. IEEE, 2023.

[209] Yu Pan, Yanni Hu, Yuguang Yang, Wen Fei, Jixun Yao, Heng Lu, Lei Ma, and Jianjun Zhao. Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10021–10025. IEEE, 2024.

[210] Zhichen Yuan, CL Philip Chen, Shuzhen Li, and Tong Zhang. Disentanglement network: Disentangle the emotional features from acoustic features for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11686–11690. IEEE, 2024.

[211] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, 2014.

[212] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[213] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[214] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, 2016.

[215] Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)*, pages 117–121. IEEE, 2020.

[216] Yoichi Takenaka. Performance evaluation of emotion classification in japanese using roberta and deberta. *arXiv preprint arXiv:2505.00013*, 2025.

[217] Kentaro Takenaka, Takumi Ito, and Daisuke Kawahara. Performance evaluation of emotion classification in japanese using roberta and deberta. *arXiv preprint arXiv:2505.00013*, 2025.

[218] Jing Zhang, Wei Li, Xin Chen, and Shulin Chang. Leveraging comet as pretrained commonsense features for emotion classification. *Transactions of the ACL*, 9:400–415, 2021.

[219] Swapna Mol George and P Muhamed Ilyas. A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015, 2024.

[220] Florian Eyben, Martin Wöllmer, and Björn Schuller. Openear — introducing the munich open-source emotion and affect recognition toolkit. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009.

[221] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[222] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE access*, 7:117327–117345, 2019.

[223] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.

[224] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

[225] Aharon Satt, Shai Rozenberg, Ron Hoory, et al. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, pages 1089–1093, 2017.

[226] Zhipeng Li, Xiaofen Xing, Yuanbo Fang, Weibin Zhang, Hengsheng Fan, and Xiangmin Xu. Multi-scale temporal transformer for speech emotion recognition. *arXiv preprint arXiv:2410.00390*, 2024.

[227] Samson Akinpelu, Serestina Viriri, and Muhammad Haroon Yousaf. Swintser: An improved bilingual speech emotion recognition using shift window transformer. *Cognitive Computation*, 17(4):129, 2025.

[228] Rashedul Hasan, Meher Nigar, Nursadul Mamun, and Sayan Paul. Emoformer: A text-independent speech emotion recognition using hybrid transformer-cnn model. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 2881–2886. IEEE, 2024.

[229] Nineli Lashkarashvili, Wen Wu, Guangzhi Sun, and Philip C Woodland. Parameter efficient finetuning for speech emotion recognition and domain adaptation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10986–10990. IEEE, 2024.

[230] Yi Chang, Zhao Ren, Zixing Zhang, Xin Jing, Kun Qian, Xi Shao, Bin Hu, Tanja Schultz, and Björn W Schuller. Staa-net: A sparse and transferable adversarial attack for speech emotion recognition. *IEEE Transactions on Affective Computing*, 2024.

[231] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[232] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023.

[233] Abinay Reddy Naini, Mary A Kohler, Elizabeth Richerson, Donita Robinson, and Carlos Busso. Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12031–12035. IEEE, 2024.

[234] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[235] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

[236] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354, 2005.

[237] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[238] Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. Emo2vec: Learning generalized emotion representation by multi-task training. *arXiv preprint arXiv:1809.04505*, 2018.

[239] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. Association for Computational Linguistics, 2014.

[240] Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume*, pages 2415–2428, 2021.

[241] Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa, and Amena Mahmoud. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, 2022(1):2645381, 2022.

[242] Hani Al-Omari, Malak A Abdullah, and Samira Shaikh. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232. IEEE, 2020.

[243] Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. Emotionx-idea: Emotion bert–an affectional model for conversation. *arXiv preprint arXiv:1908.06264*, 2019.

[244] Rohan Kamath, Arpan Ghoshal, Sivaraman Eswaran, and Prasad Honnavalli. An enhanced context-based emotion detection model using roberta. In *2022 IEEE international conference on electronics, computing and communication technologies (CONECCT)*, pages 1–6. IEEE, 2022.

[245] Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W Schuller. Refashioning emotion recognition modeling: the advent of generalized large models. *IEEE Transactions on Computational Social Systems*, 11(5):6690–6704, 2024.

[246] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, 2023.

[247] Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2202–2218, 2025.

[248] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024.

[249] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[250] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

[251] Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788, 2023.

[252] Jiajun He, Xiaohan Shi, Xingfeng Li, and Tomoki Toda. Mf-AED-AEC: Speech emotion recognition by leveraging multimodal fusion, ASR error detection, and ASR error correction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11066–11070. IEEE, 2024.

[253] Jiajun He, Jinyi Mi, and Tomoki Toda. GIA-MIC: Multimodal Emotion Recognition with Gated Interactive Attention and Modality-Invariant Learning Constraints. In *Interspeech 2025*, pages 2695–2699, 2025. doi: 10.21437/Interspeech.2025-2696.

[254] Odysseas S Chlapanis, Georgios Paraskevopoulos, and Alexandros Potamianos. Adapted multimodal bert with layer-wise fusion for sentiment analysis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[255] Zihan Zhao, Yanfeng Wang, and Yu Wang. Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In *Proc. Interspeech 2022*, pages 4725–4729, 2022.

[256] Junyi Xiang, Xianxun Zhu, and Erik Cambria. Integrating audio–visual text generation with contrastive learning for enhanced multimodal emotion analysis. *Information Fusion*, page 103809, 2025.

[257] Qifei Li, Yingming Gao, and Ya Li. Mining high-quality samples from raw data and majority voting method for multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9546–9550, 2023.

[258] Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):805–822, 2023.

[259] Ting Zhang, Bin Song, Zhiyong Zhang, and Yajuan Zhang. Multimodal sentiment analysis based on multi-stage graph fusion networks under random missing modality conditions. *IET Image Processing*, 19(1):e13310, 2025.

[260] Judith Nkechinyere Njoku, Angela C Caliwag, Wansu Lim, Sangho Kim, H Hwang, and J Jung. Deep learning based data fusion methods for multimodal emotion recognition. *The Journal of Korean Institute of Communications and Information Sciences*, 47(1):79–87, 2022.

[261] Mixiao Hou, Zheng Zhang, Chang Liu, and Guangming Lu. Semantic alignment network for multi-modal emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5318–5329, 2023.

[262] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017.

[263] Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 82–95, 2024.

[264] Rui Wang, Chaopeng Guo, Mohammad Shabaz, Imad Rida, Erik Cambria, and Xianxun Zhu. Cime: Contextual interaction-based multimodal emotion analysis with enhanced semantic information. *IEEE Transactions on Computational Social Systems*, 2025.

[265] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1726–1736, 2024.

[266] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[267] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2021.

[268] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2022.

[269] Peng He, Jun Yu, Chengjie Ge, Ye Yu, Wei Xu, Lei Wang, Tianyu Liu, and Zhen Kan. Domain-separated bottleneck attention fusion framework for multimodal emotion recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(4):1–21, 2025.

[270] Yuntao Shou, Tao Meng, Wei Ai, Fangze Fu, Nan Yin, and Keqin Li. A comprehensive survey on multi-modal conversational emotion recognition with deep learning. *ACM Transactions on Information Systems*, 2023.

[271] Yangyang Qu, Yongsheng Ou, and Rong Xiong. Low light enhancement by unsupervised network. In *2020 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 404–409. IEEE, 2020.

[272] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[273] J. Li et al. Multiemo: Correlation-aware attention for multimodal emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 16:110–125, 2024.

[274] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[275] Woo Yong Choi, Kyu Ye Song, and Chan Woo Lee. Convolutional attention networks for multimodal emotion recognition from speech and text data. In *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, pages 28–34, 2018.

[276] Zheng Lian, Bin Liu, and Jianhua Tao. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2415–2429, 2022.

[277] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161. Association for Computational Linguistics, 2018.

[278] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 852–861, 2024.

[279] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, 2022.

[280] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934, 2023.

[281] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1642–1651, 2022.

[282] Yulou Shu, Wengen Li, Yu-Ping Ruan, Wuchao Liu, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Ensuring pre-fusion modality consistency: A new approach to multimodal sentiment detection. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–20, 2025.

[283] Xiaosen Lyu, Jiayu Xiong, Yuren Chen, Wanlong Wang, Xiaoqing Dai, and Jing Wang. Cross-space synergy: A unified framework for multimodal emotion recognition in conversation. *arXiv preprint arXiv:2512.03521*, 2025.

[284] Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, K.Y. Zheng, and Qunyan Zhou. Acformer: An aligned and compact transformer for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 833–842. ACM, 2023.

[285] Tao Zhang and Zhenhua Tan. ECERC: evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2064–2077, 2025.

[286] Zechang Xiong, Zhenyan Ji, Wenkang Kong, Jiuqian Dai, and Shen Yin. Esed: Emotion-specific evidence decomposition for uncertainty-aware multimodal emotion recognition in conversation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3582–3591, 2025.

[287] Jingyao Wu, Grace Lin, Yinuo Song, and Rosalind Picard. Amber ˆ2: Dual ambiguity-aware emotion recognition applied to speech and text. *arXiv preprint arXiv:2601.18010*, 2026.

[288] Guowei Zhong, Ruohong Huan, Mingzhen Wu, Ronghua Liang, and Peng Chen. Towards robust multimodal emotion recognition under missing modalities and distribution shifts. *arXiv preprint arXiv:2506.10452*, 2025.

[289] Wen Yin, Siyu Zhan, Cencen Liu, Xin Hu, Guiduo Duan, Xiurui Xie, Yuan-Fang Li, and Tao He. Tical: Typicality-based consistency-aware learning for multimodal emotion recognition. *arXiv preprint arXiv:2511.15085*, 2025.

[290] Shuai Liu, Peng Gao, Yating Li, Weina Fu, and Weiping Ding. Multi-modal fusion network with complementarity and importance for emotion recognition. *Information Sciences*, 619:679–694, 2023.

[291] Changzeng Fu, Fengkui Qian, Kaifeng Su, Yikai Su, Ze Wang, Jiaqi Shi, Zhigang Liu, Chaoran Liu, and Carlos Toshinori Ishi. Himul-lgg: A hierarchical decision fusion-based local–global graph neural network for multimodal emotion recognition in conversation. *Neural Networks*, 181:106764, 2025.

[292] Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. Dynamic graph neural ode network for multi-modal emotion recognition in conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 256–268, 2025.

[293] Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11418–11426, 2025.

[294] Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, and Heikki Kälviäinen. Deemo: De-identity multimodal emotion recognition and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5707–5716, 2025.

[295] Johannes Wagner, Elisabeth Andre, Florian Lingenfelser, and Jonghwa Kim. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218, 2011.

[296] Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

[297] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101:101973, 2024.

[298] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1545–1554, 2022.

[299] Xianxun Zhu, Yaoyang Wang, Erik Cambria, Imad Rida, José Santamaría López, Lin Cui, and Rui Wang. Rmer-dt: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. *Information Fusion*, page 103268, 2025.

[300] Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, 261:217–230, 2017.

[301] Junyin Peng, Hong Tang, and Wenbin Zheng. Hierarchical heterogeneous graph network based multimodal emotion recognition in conversation. *Multimedia Systems*, 31(2):81, 2025.

[302] Yijing Dai, Yingjian Li, Dongpeng Chen, Jinxing Li, and Guangming Lu. Multimodal decoupled distillation graph neural network for emotion recognition in conversation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9910–9924, 2024.

[303] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.

[304] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[305] Haiping Huang, Zhenchao Hu, Wenming Wang, and Min Wu. Multimodal emotion recognition based on ensemble convolutional neural network. *IEEE Access*, 8: 3265–3271, 2019.

[306] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 529–535, 2017.

[307] Enguerrand Boitel, Alaa Mohasseb, and Ella Haig. Mist: Multimodal emotion recognition using deberta for text, semi-cnn for speech, resnet-50 for facial, and 3d-cnn for motion analysis. *Expert Systems with Applications*, 270:126236, 2025.

[308] Cheng Cheng, Wenzhe Liu, Xinying Wang, Lin Feng, and Ziyu Jia. Disd-net: A dynamic interactive network with self-distillation for cross-subject multi-modal emotion recognition. *IEEE Transactions on Multimedia*, 2025.

[309] Jiaxing Liu, Sen Chen, Longbiao Wang, Zhilei Liu, Yahui Fu, Lili Guo, and Jianwu Dang. Multimodal emotion recognition with capsule graph convolutional based representation fusion. In *ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 6339–6343. IEEE, 2021.

[310] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026, 2020.

[311] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. In *Proc. Interspeech 2019*, pages 3569–3573, 2019.

[312] Rory Beard, Ritwik Das, Raymond WM Ng, PG Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik. Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proceedings of the 22nd conference on computational natural language learning*, pages 251–259, 2018.

[313] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020.

[314] Minh Tran and Mohammad Soleymani. A pre-trained audio-visual transformer for emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4698–4702. IEEE, 2022.

[315] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382, 2024.

[316] Kamran Ali and Charles E Hughes. A unified transformer-based network for multimodal emotion recognition. *arXiv preprint arXiv:2308.14160*, 2023.

[317] Dung Nguyen, Duc Thanh Nguyen, Rui Zeng, Thanh Thi Nguyen, Son N Tran, Thin Nguyen, Sridha Sridharan, and Clinton Fookes. Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Transactions on Multimedia*, 24:1313–1324, 2021.

[318] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

[319] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4703–4707. IEEE, 2022.

[320] Zexu Pan, Zhaojie Luo, Jichen Yang, and Haizhou Li. Multi-modal attention for speech emotion recognition. *arXiv preprint arXiv:2009.04107*, 2020.

[321] Ngoc-Huynh Ho, Hyung-Jeong Yang, Soo-Hyung Kim, and Gueesang Lee. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, 8:61672–61686, 2020.

[322] Ziwang Fu, Feng Liu, Hanyang Wang, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li. A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv preprint arXiv:2111.02172*, 2021.

[323] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279. IEEE, 2021.

[324] Tianyi Wu, Erick Purwanto, Yongrun Huang, and Su Yang. Phy-fusionnet: A memory-augmented transformer for multimodal emotion recognition with periodicity and contextual attention. *IEEE Transactions on Affective Computing*, 2025.

[325] Zihan Zhao, Yu Wang, and Yanfeng Wang. Knowledge-aware bayesian co-attention for multimodal emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[326] Bubai Maji, Monorama Swain, Rajlakshmi Guha, and Aurobinda Routray. Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[327] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2486–2495, 2022.

[328] Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, and Soonil Kwon. Mser: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, 245:122946, 2024.

[329] Zebang Cheng, Shuyuan Tu, Dawei Huang, Minghan Li, Xiaojiang Peng, Zhi-Qi Cheng, and Alexander G Hauptmann. Sztu-cmu at mer2024: Improving emotion-llama with conv-attention for multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 78–87, 2024.

[330] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017.

[331] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *Ieee Access*, 8:176274–176285, 2020.

[332] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors*, 21(14):4913, 2021.

[333] Yuhua Wang, Guang Shen, Yuezhu Xu, Jiahang Li, and Zhengdao Zhao. Learning mutual correlation in multimodal transformer for speech emotion recognition. In *Interspeech*, pages 4518–4522. Cary, NC, 2021.

[334] Rutherford Agbeshi Patamia, Wu Jin, Kingsley Nketia Acheampong, Kwabena Sarpong, and Edwin Kwadwo Tenagyei. Transformer based multimodal speech emotion recognition with improved neural networks. In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 195–203. IEEE, 2021.

[335] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437, 2022.

[336] Feng Zhang, Xi-Cheng Li, Chee Peng Lim, Qiang Hua, Chun-Ru Dong, and Jun-Hai Zhai. Deep emotional arousal network for multimodal sentiment analysis and emotion recognition. *Information Fusion*, 88:296–304, 2022.

[337] ShiHao Zou, Xianying Huang, XuDong Shen, and Hankai Liu. Improving multimodal fusion with main modal transformer for emotion recognition in conversation. *Knowledge-Based Systems*, 258:109978, 2022.

[338] Ye Jing and Xinpei Zhao. Dq-former: Querying transformer with dynamic modality priority for cognitive-aligned multimodal emotion recognition in conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4795–4804, 2024.

[339] Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *proceedings of the 31st ACM international conference on multimedia*, pages 5975–5984, 2023.

[340] Hoai-Duy Le, Guee-Sang Lee, Soo-Hyung Kim, Seungwon Kim, and Hyung-Jeong Yang. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *Ieee Access*, 11:14742–14751, 2023.

[341] Jia-Hao Hsu and Chung-Hsien Wu. Applying segment-level attention on bi-modal transformer encoder for audio-visual emotion recognition. *IEEE Transactions on Affective Computing*, 14(4):3231–3243, 2023.

[342] Shihao Zou, Xianying Huang, and Xudong Shen. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5994–6003, 2023.

[343] Zuojin Tang, Bin Hu, Chenyang Zhao, De Ma, Gang Pan, and Bin Liu. Vlascd: A visual language action model for simultaneous chatting and decision making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9223–9243, 2025.

[344] Chengxin Chen and Pengyuan Zhang. Modality-collaborative transformer with hybrid feature reconstruction for robust emotion recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5):1–23, 2024.

[345] Hajar Filali, Chafik Boulealam, Khalid El Fazazy, Adnane Mohamed Mahraz, Hamid Tairi, and Jamal Riffi. Meaningful multimodal emotion recognition based on capsule graph transformer architecture. *Information*, 16(1):40, 2025.

[346] Feng Liu, Ziwang Fu, Yunlong Wang, and Qijian Zheng. Tacfn: transformer-based adaptive cross-modal fusion network for multimodal emotion recognition. *arXiv preprint arXiv:2505.06536*, 2025.

[347] Fei Ma, Yang Li, Shiguang Ni, Shao-Lun Huang, and Lin Zhang. Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional gan. *Applied Sciences*, 12(1):527, 2022.

[348] Minjie Ren, Xiangdong Huang, Jing Liu, Ming Liu, Xuanya Li, and An-An Liu. Maln: Multimodal adversarial learning network for conversational emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11):6965–6980, 2023.

[349] Jiajun He, Xiaohan Shi, Cheng-Hung Hu, Jinyi Mi, Xingfeng Li, and Tomoki Toda. M^{4}SER: Multimodal, multirepresentation, multitask, and multistrategy learning for speech emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4055–4070, 2025.

[350] Cheng Cheng, Wenzhe Liu, Zhaoxin Fan, Lin Feng, and Ziyu Jia. A novel transformer autoencoder for multi-modal emotion recognition with incomplete data. *Neural Networks*, 172:106111, 2024.

[351] Wenjin Tian, Xianying Huang, and Shihao Zou. Multi-condition guided diffusion network for multimodal emotion recognition in conversation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3215–3227, 2025.

[352] Xiongjian Lv, Yimin Wen, and Hang Yu. Diffufuse: Diffusion-driven dual-stream fusion framework for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8458–8467, 2025.

[353] Cheng Cheng, Wenzhe Liu, Yong Zhang, Lin Feng, and Ziyu Jia. A cross-modal adaptive masked autoencoder for decoding emotions with multimodal data. *IEEE Transactions on Computational Social Systems*, 2024.

[354] Yuehan Jin, Xiaoqing Liu, Yiyuan Yang, Zhiwen Yu, Tong Zhang, and Kaixiang Yang. Rohydr: Robust hybrid diffusion recovery for incomplete multimodal emotion recognition. *arXiv preprint arXiv:2505.17501*, 2025.

[355] Haoqin Sun, Xugang Lu, Jingguang Tian, Jiaming Zhou, Jiabei He, Hui Wang, Xiangyu Kong, Xinhui Hu, and Yong Qin. Progressive learning framework with missing modality reconstruction for multimodal emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 34:29–43, 2025.

[356] Yuanyuan Sun and Ting Zhou. Dialoguemllm: transforming multimodal emotion recognition in conversation through instruction-tuned mllm. *IEEE Access*, 2025.

[357] Chenyu Zhang, Minsol Kim, Shohreh Ghorbani, Jingyao Wu, Rosalind Picard, Patricia Maes, and Paul Pu Liang. When one modality sabotages the others: A diagnostic lens on multimodal reasoning. *arXiv preprint arXiv:2511.02794*, 2025.

[358] Keane Ong, Wei Dai, Carol Li, Dewei Feng, Hengzhi Li, Jingyao Wu, Jiaee Cheong, Rui Mao, Gianmarco Mengaldo, Erik Cambria, et al. Human behavior atlas: Benchmarking unified psychological and social behavior understanding. *arXiv preprint arXiv:2510.04899*, 2025.

[359] Keane Ong, Sabri Boughorbel, Luwei Xiao, Chanakya Ekbote, Wei Dai, Ao Qu, Jingyao Wu, Rui Mao, Ehsan Hoque, Erik Cambria, et al. Omnisapiens: A foundation model for social behavior processing via heterogeneity-aware relative policy optimization. *arXiv preprint arXiv:2602.10635*, 2026.

[360] Yuanchao Li, Yuan Gong, Chao-Han Huck Yang, Peter Bell, and Catherine Lai. Revise, reason, and recognize: Llm-based emotion recognition via emotion-specific prompts and asr error correction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[361] Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, and Xiaojiang Peng. Emotion-qwen: Training hybrid experts for unified emotion and general vision-language understanding. 2025.

[362] Bohao Xing, Xin Liu, Guoying Zhao, Chengyu Liu, Xiaolan Fu, and Heikki Kälviäinen. Emotionhallucer: Evaluating emotion hallucinations in multimodal large language models. *arXiv preprint arXiv:2505.11405*, 2025.
[363] Laurence Devillers and Roddy Cowie. Ethical considerations on affective computing: An overview. *Proceedings of the IEEE*, 111(10):1445–1458, 2023.
[364] Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, and Xie Chen. Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11146–11150. IEEE, 2024.
[365] Huan Liu, Xin Zhang, Junyang Wei, et al. How to bridge the gap between modalities: Survey on multimodal large language models. *arXiv preprint arXiv:2501.00001*, 2025.
[366] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. *The Oxford handbook of affective computing*. Oxford University Press, 2015.
[367] Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638*, 2024.