Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects

Chengyan Wu[*1,2], Yiqiang Cai[*1,2], Yang Liu[3], Pengxu Zhu[4], Yun Xue[‡1,2], Ziwei Gong[5], Julia Hirschberg[5], Bolei Ma[‡6]

[1]Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials

[2]School of Electronic Science and Engineering (School of Microelectronics), South China Normal University

[3]North Carolina Central University [4]Georgia Institute of Technology [5]Columbia University

[6]LMU Munich & Munich Center for Machine Learning

[*]Equal contributions. [‡]Corresponding authors.

{chengyan.wu, yiqiangcai, xueyun}@m.scnu.edu.cn, zg2272@columbia.edu, bolei.ma@lmu.de

Abstract

While text-based emotion recognition methods have achieved notable success, real-world dialogue systems often demand a more nuanced emotional understanding than any single modality can offer. Multimodal Emotion Recognition in Conversations (MERC) has thus emerged as a crucial direction for enhancing the naturalness and emotional understanding of human-computer interaction. Its goal is to accurately recognize emotions by integrating information from various modalities such as text, speech, and visual signals.

This survey offers a systematic overview of MERC, including its motivations, core tasks, representative methods, and evaluation strategies. We further examine recent trends, highlight key challenges, and outline future directions. As interest in emotionally intelligent systems grows, this survey provides timely guidance for advancing MERC research.

1 Introduction

Emotion recognition in conversations (ERC) (Peng et al., 2022; Deng and Ren, 2023) is an increasingly important task in natural language processing (NLP), focusing on identifying the emotional state associated with each utterance in a dialogue. Unlike conventional emotion classification on isolated sentences, ERC requires understanding the interplay between utterances and tracking speaker-specific context across the conversation (Gao et al., 2024). Its relevance has grown due to its potential in various real-world applications, including social media monitoring (Kumar et al., 2015), intelligent healthcare services (Hu et al., 2021b), and the design of emotionally aware dialogue agents (Jiao et al., 2020; Gong et al., 2023).

However, human emotions are typically conveyed through multiple modalities, including auditory (e.g., speech), visual (e.g., facial expressions, gestures), and linguistic (e.g., the semantic content conveyed by transcribed text). As a result, recent research (e.g., Ma et al., 2024a; Van et al., 2025; Dutta and Ganapathy, 2025) has increasingly focused on multimodal settings in dialogue, a direction we refer to as Multimodal Emotion Recognition in Conversations (MERC). Researchers aim to identify the emotional state of a given utterance by integrating contextual information from different modalities, which often includes subtle personal emotional states such as happiness, anger, and hatred (Poria et al., 2019b; Gong et al., 2024), thereby improving the effectiveness of emotion recognition in dialogues. Figure 1 illustrates an example of ERC with textual, acoustic, and visual inputs.

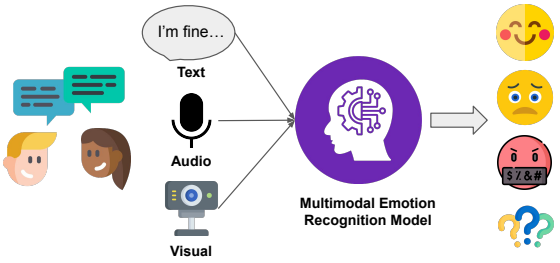

Figure 1: An example of MERC. Text, audio, and visual inputs are integrated through a multimodal model to detect various emotional states.

Multimodal emotion recognition (MER) itself has gained increasing attention due to the challenges of integrating diverse modalities, prompting research in both non-conversational and conversational settings. Existing surveys such as A.V. et al. (2024) and Aruna Gladys and Vetriselvi (2023) focus on non-conversational MER but overlook key aspects like interlocutor modeling and context. Fu et al. (2023) reviews both unimodal and multimodal conversational MER, yet primarily centers on fea-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 6257–6274

ture fusion, offering limited insight into core challenges such as cross-modal alignment, reasoning, modality missingness, and conflicts.

Despite growing interest, the MERC task remains underexplored. Existing surveys (Fu et al., 2023; Zhang and Tan, 2024) also lag behind recent advances, particularly the rise of multimodal large language models (MLLMs). To bridge this gap, we present a timely and comprehensive review of MERC. We first introduce the task definition and our survey methodology (§2), benchmark datasets and evaluation methods (§3), followed by a review of preprocessing techniques (§4); then categorize recent methods (§5) and outline key challenges and prospects (§6).

In summary, the specific contributions of this survey are threefold:

- **Compilation of recent MERC research developments.** We systematically review and integrate the latest research progress made by MERC in recent years, covering diverse datasets and methodologies.
- **Summarizing and comparing various MERC methods.** We evaluate the strengths and limitations of various MERC approaches, offering theoretical insights and practical guidance to help researchers and practitioners select appropriate methods.
- **Proposing challenges and future directions.** We identify key open issues in the MERC domain and put forward several potential future research directions, aiming to guide ongoing and future investigations by researchers and practitioners in MERC.

2 Task Settings and Review Methodology

In this section, we present the task settings of MERC and outline the methodology employed to compile the content of this survey, which details the strategy and selection criteria used to curate the final content for this survey paper.

Modalities. In the context of MERC, we define a modality as a distinct source or channel of information that conveys emotional or communicative signals. These modalities correspond to observable representations derived from different sensory inputs and are typically processed in separate feature spaces. The most common modalities include:

- **Textual modality**: The transcribed representation of spoken utterances, capturing the semantic and syntactic content of language.
- **Acoustic modality**: Prosodic and paralinguistic features extracted from speech, such as tone, pitch, and energy.
- **Visual modality**: Non-verbal cues such as facial expressions, head movements, eye gaze, and gestures.

Task Definition. Given a dialogue $D = \{u_1, u_2, \ldots, u_N\}$ consisting of N utterances spoken by multiple speakers, the goal of the MERC task is to predict an emotion label $e_i \in \mathcal{Y}$ for each utterance u_i. Each utterance is associated with three modalities: textual (u_i^t), acoustic (u_i^a), and visual (u_i^v), which provide complementary information for emotion recognition. The multimodal representation of an utterance is denoted as:

$$u_i = [u_i^t; u_i^a; u_i^v], \quad \text{for } i = 1, 2, \ldots, N \quad (1)$$

where $[\cdot;\cdot;\cdot]$ denotes combination of modalities.

Methodology for Literature Compilation:

Strategy. We conduct a comprehensive literature search using sources such as the ACL Anthology, Google Scholar, and general search engines (e.g., Google Chrome). Within the ACL Anthology, we focus on top venues including EMNLP, ACL, NAACL, and related workshops.[1]

Selection Criteria. We select papers that are directly relevant to MERC, with a focus on works that used at least two modalities (e.g., text, audio, visual), included conversational context, and evaluated on benchmark datasets such as IEMOCAP, MELD, and CMU-MOSEI. We prioritize recent papers from 2020 onward to reflect the state-of-the-art, while including foundational work when appropriate for historical context. The selection is based on a careful review of the abstract, introduction, conclusion, and limitations of each paper.

3 Datasets and Evaluation

In this section, we describe the evaluation datasets and the evaluation metrics used for the MERC task, focusing on multimodal resources across multiple

[1]We used keywords such as *"multimodal emotion recognition", "emotion recognition in conversation", "multimodal affective computing", "dialogue emotion recognition", "MERC benchmark", "context-aware emotion detection",* and *"multimodal conversational modeling"*.

languages. For more detailed information about the single benchmarks, see Appendix §A.

(1) *English-centered:* IEMOCAP, MELD, CMU-MOSEI, AVEC, EmoryNLP, and MEmoR dataset.

(2) *Non-English:* M-MELD (French, Spanish, Greek, Polish), ACE (African), M^{3}ED (Mandarin).

As shown in Table 1, the domains covered by multimodal datasets have become increasingly diverse over time. The sources of these datasets include TV series, videos, and movies. At the same time, linguistic diversity has expanded to include languages such as French, Spanish, Greek, Polish, and Mandarin. Notably, there has also been a growing emergence of datasets targeting low-resource languages, such as African languages.

Datasets	Lang.	Source	Year
IEMOCAP (Busso et al., 2008)	en	Videos	2008
AVEC (Schuller et al., 2012)	en	Videos	2012
EmoryNLP (Zahiri and Choi, 2017)	en	TV series	2017
CMU-MOSEI (Bagher Zadeh et al., 2018)	en	Videos	2018
MELD (Poria et al., 2019a)	en	TV series	2019
MEmoR (Shen et al., 2020)	en	Videos	2020
M^{3}ED (Zhao et al., 2022)	zh	TV series	2022
M-MELD (Ghosh et al., 2023)	fr,es,el,pl	TV series	2023
ACE (Sasu et al., 2025)	Akan	Movies	2025

Table 1: Overview of Datasets.

Datasets. We divide the existing mainstream dataset into the following two categories:

Evaluation Metrics. Existing studies (Majumder et al., 2019; Ghosal et al., 2019, *inter alia*) typically adopt multiple evaluation metrics to comprehensively assess the overall performance of models, including **Accuracy** (e.g., Shou et al., 2024), **Weighted-F1** (e.g., Ma et al., 2024a), **Macro-F1** (e.g., Chudasama et al., 2022), and **Micro-F1** (e.g., Xie et al., 2021) scores. To enable fine-grained analysis, these works also report per-emotion metric scores.

4 Feature Processing

Preprocessing dataset features is essential for effectively extracting meaningful information. We summarize feature preprocessing methods used in prior MERC research and analyze the typical preprocessing pipeline, which is often tailored to conversational settings. Specifically, we distinguish between two key components: **feature extraction** and **context modeling**.

4.1 Feature Extraction

For effective multimodal analysis, features must first be extracted from each modality stream (text, audio, visual). Mainstream approaches (e.g., Shi and Huang, 2023) typically process these modalities separately at this initial stage. While the core extraction techniques often overlap, the key difference in the multimodal setting lies in the objective and subsequent use of these features. In unimodal ERC, the extractor aims to capture enough information within that single modality for emotion classification. Table 2 provides an overview of common feature extraction models employed in the multimodal studies surveyed in this paper.

Modality	Models
Text	LSTM (Hochreiter and Schmidhuber, 1997) CNN (Kim, 2014; Tran et al., 2015) Transformer (Vaswani et al., 2017) RoBERTa (Liu et al., 2019) sBERT (Reimers and Gurevych, 2019)
Visual	3D-CNN (Tran et al., 2015) OpenFace (Baltrušaitis et al., 2016) MTCNN (Zhang et al., 2016) DenseNet (Huang et al., 2017) VisExtNet (Shi and Huang, 2023)
Audio	openSMILE (Eyben et al., 2010) COVAREP (Degottex et al., 2014) librosa (McFee et al., 2015) DialogueRNN (Majumder et al., 2019)

Table 2: Feature Extraction models.

4.2 Context Modeling

Context modeling primarily involves two types of contextual dependencies: **situation-level** and **speaker-level** modeling.

Situation-level. The emotional state of a speaker is influenced not only by the semantic content of the current utterance but also by the surrounding contextual semantics. Therefore, existing approaches (Hu et al., 2021a; Majumder et al., 2019) commonly employ specialized networks to model the sequential dependencies among utterances, aiming to more accurately capture the speaker's temporal emotional dynamics. Given the textual feature $u_i \in \mathbb{R}^{d_u}$ for each utterance, the sequential context representation $c_i^s \in \mathbb{R}^{2d_u}$ is computed as follows:

$$\mathbf{c}_i^s,\ \mathbf{h}_i^s = \text{Model}\left(\mathbf{u}_i,\ \mathbf{h}_{i-1}^s\right) \quad (2)$$

where $h_i^s \in \mathbb{R}^{d_u}$ represents the i-th hidden state of the contextual sequence.

Speaker-level. Speaker identity information typically exhibits temporal and relational properties of emotions, which can enhance the model's ability

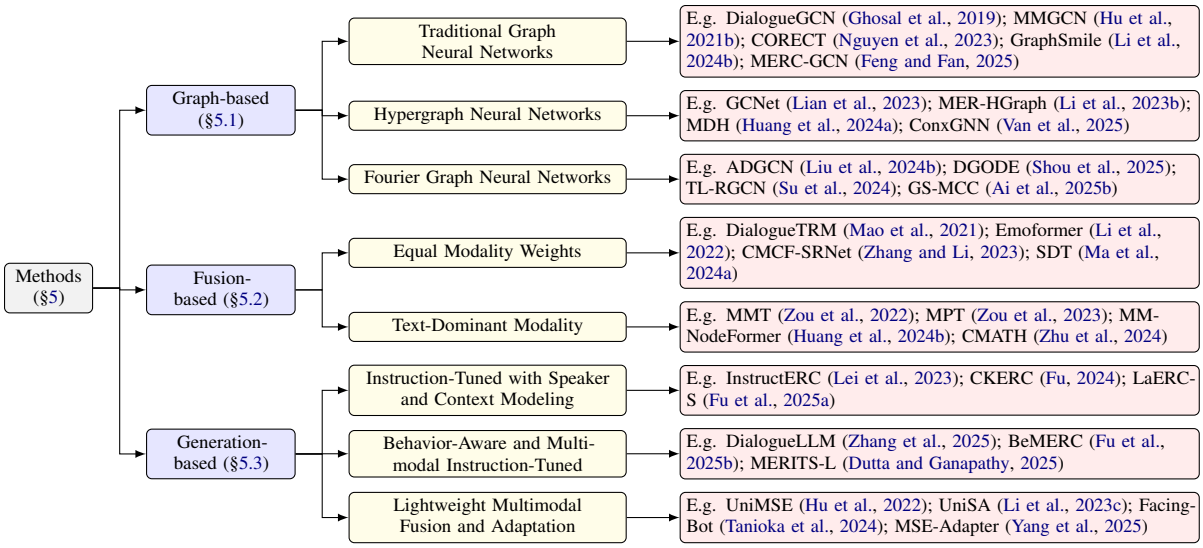

Figure 2: A taxonomy of mainstream methods.

to perceive speaker role information. Therefore, to more effectively learn and distinguish speaker-level contextual representations, many studies have further introduced speaker-related structured information on top of dialogue context modeling. Common approaches include using ***Speaker Embeddings*** (Ma et al., 2024a; Shen et al., 2020) to explicitly differentiate between different speakers, or leveraging ***Graph Neural Networks*** (Ai et al., 2025a; Van et al., 2025) to construct interaction graphs between speakers, thereby more comprehensively modeling the dependencies between them:

Speaker Embeddings: The speaker embedding S_i is combined with the modality features (e.g., text, audio, or vision) to produce modality representations that are speaker- and context-aware.

$$\mathbf{x}_m = \mathbf{c}_i^s + \mathbf{S}_i, \quad m \in \{t, a, v\} \tag{3}$$

Graph Neural Networks: Speaker interactions can be further modeled by constructing a dialogue graph to capture inter-utterance and inter-speaker dependencies beyond sequential semantics. A dialogue graph is typically defined as $G = (V, E, W, R)$, where each node $v_i \in V$ corresponds to the i-th utterance, and its associated feature vector $\mathbf{c}_i^s$ is obtained from sequential modeling of contextual dependencies. Edges $e_{ij} \in E$ represent interaction links between utterances, with associated weights $\omega_{ij} \in W$ reflecting the interaction strength and types $r_{ij} \in R$ encoding speaker-related or structural relationships.

Based on this graph, the context-aware representation $\mathbf{h}_i^g$ for node v_i is computed using a graph neural network as follows:

$$\mathbf{h}_i^g = \mathrm{GNN}(\mathbf{c}_i^s, \{(\mathbf{c}_j^s, \omega_{ij}, r_{ij}) \mid e_{ij} \in E\}) \tag{4}$$

where $\mathbf{c}_j^s$ are the contextual node features from neighboring utterances. The GNN aggregates information from connected nodes to enhance $\mathbf{c}_i^s$ with speaker- and structure-aware interaction knowledge.

5 Methodology

This section discusses state-of-the-art approaches to the MERC tasks. We summarize them from three perspectives: Graph-based Methods (§5.1), Fusion-based Methods (§5.2), and Generation-based Methods (§5.3). An overview of the methods and subcategories with representative examples is presented in Figure 2.

It is worth noticing that some methods in MERC inherently involve multiple components (e.g., fusion modules within graph-based or generation-based frameworks). Our taxonomy is not intended to be strictly disjoint, but rather to organize the literature based on the core modeling paradigm or innovation focus of each method. The categorization of methods is thus as follows:

- **Graph-based**: Methods are categorized as graph-based when the primary architecture centers around graph neural networks for modeling conversational structure, even if auxiliary modules (e.g., fusion layers) are integrated.

- **Fusion-based**: Methods are grouped under fusion-based when their main contribution lies in the design of cross-modal interaction mechanisms, regardless of the backbone architecture (e.g., Transformer, LSTM).

- **Generation-based**: Generation-based methods refer to recent approaches that leverage LLMs to generate predictions or intermediate reasoning, often using prompt engineering or instruction tuning, even if lightweight fusion components are present.

5.1 Graph-based Methods

Dialogues can be naturally interpreted as graph structures due to the intrinsic correlations and dependencies among utterances. Conversations often feature multi-turn interactions with complex dependency and interaction patterns, which can be effectively modeled through the edge structures of Graph Neural Networks (GNNs) (Scarselli et al., 2009). With the increasing interest in multimodal dialogue understanding, GNNs have evolved beyond their application (Liu et al., 2024a) in textual data to embrace multimodal inputs. Besides, recent methods also integrate auxiliary modules (e.g., convolution, contrastive learning, and fusion) to enhance the performance. Figure 3 illustrates recent advancements in graph-based methods. We categorize them into traditional graphs, hypergraphs, and fourier graph neural networks.

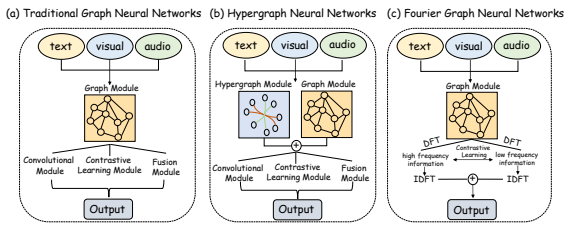

Figure 3: Development of Graph-based Methods.

Traditional Graph Neural Networks. Previous works such as bc-LSTM (Poria et al., 2017) and ICON (Hazarika et al., 2018) primarily relied on sequential approaches. DialogueGCN (Ghosal et al., 2019) was the first to introduce GNNs into ERC, addressing the limitations of earlier sequence-based models like DialogueRNN (Majumder et al., 2019) in capturing contextual dependencies. To effectively integrate information from different modalities, Hu et al. (2021b) constructed a graph structure that fuses multimodal features, enabling the model to capture inter-modal dependencies through graph convolutional networks and incorporating speaker information to enhance the representation of conversational semantics. Inspired by the application of graph convolutions in ERC, Li et al. (2024b) proposed the GSF module, which introduces an alternating graph convolution mechanism to hierarchically extract both cross-modal and intra-modal emotional information. Some studies further enhanced graph-based models with attention mechanisms for multimodal fusion; for example, Feng and Fan (2025) integrated a Cross-modal Attention Module to better fuse information from different modalities, while Nguyen et al. (2023) designed a cross-modal attention mechanism to explicitly model the heterogeneity between modalities.

Hypergraph Neural Networks. Although traditional graph-based methods can capture long-range and multimodal contextual information, they are often challenged by missing modalities during conversations. Lian et al. (2023) tackled this issue by jointly optimizing classification and reconstruction tasks in an end-to-end manner to effectively model incomplete data. The related works (Li et al., 2023b; Huang et al., 2024a) considered the limitations imposed by the pairwise relationships between GNNs nodes.Van et al. (2025) constructed a multimodal fusion graph and introduced Hypergraph Neural Networks (Feng et al., 2019) to connect multiple modalities or utterance nodes simultaneously, thereby capturing more complex multivariate dependencies and high-order interactions in conversations, and enhancing the modeling of emotion propagation.

Fourier Graph Neural Networks. Increasing the depth of GNN layers can lead to the oversmoothing problem (Liu et al., 2022; Yi et al., 2023), which hampers the modeling of long-range semantic dependencies and complementary modality relations. To address this issue, GS-MMC (Ai et al., 2025b) proposes a graph-based framework for multimodal consistency and complementarity learning. This method employs a Fourier graph operator to extract high- and low-frequency emotional signals from the frequency domain, capturing both local variations and global semantic trends. Additionally, a contrastive learning mechanism (van den Oord et al., 2019) is designed to enhance the semantic consistency and complementarity of these signals in a self-supervised manner, thereby improving the model's ability to recognize true emotional states.

Graph-based methods effectively capture long-

range dependencies and speaker interactions by modeling utterances as nodes and relations as edges. In traditional NLP tasks, they are particularly effective for structured inputs such as token-level representations (Zhang et al., 2023, 2024b). However, integrating heterogeneous multimodal signals into graph structures remains challenging, as naïve connections may introduce noise without proper modality alignment.

5.2 Fusion-based Methods

In MERC, effective fusion of heterogeneous multimodal features is crucial but challenging due to noise introduced during interaction modeling. The advancements of the Transformers architecture (Vaswani et al., 2017), with its self-attention mechanism, promote advancements of the MERC methods for capturing cross-modal and contextual dependencies. To enhance cross-modal interactions, recent methods build on Transformers with tailored fusion strategies. We refer to these methods as fusion-based and illustrate them in Figure 4. Some approaches promote equal interaction among modalities to improve robustness, while others adopt a primary-auxiliary scheme, typically using text as the core, with other modalities providing complementary signals.

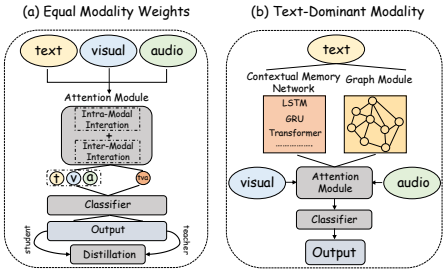

Figure 4: Development of Fusion-based Methods.

Equal Modality Weights. Equal interaction can fully utilize information from various modalities, preventing over-reliance on a single modality. Li et al. (2022) proposed achieving emotion recognition by equally integrating emotional vectors and sentence vectors from different modalities to form emotion capsules. Zhang and Li (2023) designed a local constraint module for modalities within the Transformer to promote modality interaction and incorporates a semantic graph to address the lack of semantic relationship information between utterances. Mao et al. (2021) constructed a hierarchical Transformer where each modality can flexibly switch between sequential and feedforward structures based on contextual information. Inspired by hierarchical modality interaction, Ma et al. (2024a) introduced a hierarchical gating (Ma et al., 2019) fusion strategy into the Transformer architecture to enable fine-grained modality interaction and designs self-distillation (Zhang et al., 2019) to further learn better modality representations.

Text-Dominant Modality. Some methods propose models based on primary–auxiliary modality collaboration, where auxiliary modalities are used to enhance the performance of the primary (textual) modality. Huang et al. (2024b) suggested that enhancing a text-dominant model with auxiliary modalities can improve performance. Zou et al. (2022) employed a Transformer architecture to design cross-modal attention for learning fusion relationships between different modalities, preserving the integrity of the primary modality's features while enhancing the representation of weaker modality features. It also uses a two-stage emotional cue extractor to extract emotional evidence. Building on this, Zou et al. (2023) proposed using weaker modalities as multimodal prompts while performing deep emotional cue extraction for stronger modalities. The cue information is embedded into various attention layers of the Transformer to facilitate the fusion of information between the primary and auxiliary modalities. Zhu et al. (2024) introduced an asymmetric CMA-Transformer module for central and auxiliary modalities to obtain fused modality information and proposes a hierarchical distillation (Yang et al., 2021) framework to perform coarse- and fine-grained distillation. This approach ensures the consistency of modality fusion information at different granularities.

Fusion-based methods focus on learning cross-modal interactions through attention mechanisms, with Transformer-based models achieving strong generalization. These approaches are efficient for tasks with well-aligned modality inputs but often overlook dialogue-level structures such as speaker dependencies. Compared to graph-based models, they emphasize modality-level fusion over relational reasoning.

5.3 Generation-based Methods

In recent years, pretrained LLMs have achieved remarkable success in natural language processing tasks (Chu et al., 2024), demonstrating strong emergent capabilities (Wei et al., 2022). However, despite their powerful general-purpose abilities, leveraging their full potential in specific sub-tasks still

requires carefully crafted, high-quality prompts (Wei et al., 2021) to bridge the gap in reasoning capabilities.As shown in Figure 5, researchers have proposed various model improvement strategies to effectively integrate contextual and multimodal information into LLMs while addressing their substantial computational resource demands.

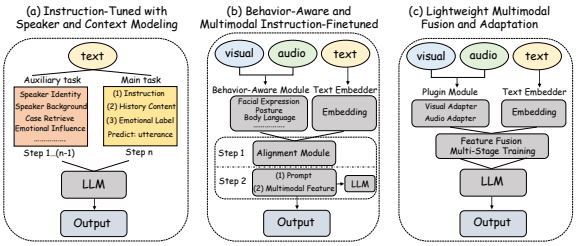

Figure 5: Development of Generation-based Methods.

Instruction-Tuned with Speaker and Context Modeling. ERC tasks have predominantly relied on discriminative modeling frameworks. With the emergence of LLMs, InstructERC (Lei et al., 2023) was the first to propose a generative framework for ERC. It introduces a simple yet effective retrieval-based prompting module that helps LLMs explicitly integrate multi-granularity supervisory signals from dialogues. Additionally, it incorporates an auxiliary emotion alignment task to better model the complex emotional transitions between interlocutors in conversations. Inspired by the integration of commonsense knowledge in COSMIC (Ghosal et al., 2020), recent work (Fu, 2024; Fu et al., 2025a) designed a prompt generation approach based on dialogue history to elicit speaker-related commonsense using LLMs by injecting commonsense knowledge into ERC.

Behavior-Aware and Multimodal Instruction-Tuned. To address the lack of multimodal integration, Dutta and Ganapathy (2025) incorporated both acoustic and textual modalities. Considering that visual information may provide richer emotional cues, Zhang et al. (2025) constructed a high-quality instruction dataset using image and text data, and fine-tunes the model using Low-Rank Adaptation (LoRA) (Song et al., 2024). Furthermore, Fu et al. (2025b) introduced a novel behavior-aware Multimodal LLM (MLLM)-based ERC framework. It consists of three core components: a video-derived behavior generation module, a behavior alignment and refinement module, and an instruction tuning module (Wei et al., 2021). The first two modules enable the model to infer human behaviors from limited information, thereby enhancing its behavioral perception capability. The instruction tuning module improves the model's emotion recognition performance by aligning and fine-tuning the concatenated multimodal inputs.

Lightweight Multimodal Fusion and Adaptation. As LLMs become increasingly large, the computational costs for ERC also rise significantly (Zhang et al., 2025). Inspired by domain-specific LLM paradigms tailored for affective computing (Hu et al., 2022; Li et al., 2023c; Tanioka et al., 2024), MSE-Adapter (Yang et al., 2025) proposed a lightweight and adaptable plug-in architecture with two modules: TGM, for aligning textual and non-textual features, and MSF, for multi-scale cross-modal fusion. Built on a frozen LLM backbone and trained via backpropagation, it enables efficient and multimodally-aware ERC with minimal computational cost. Similarly, SpeechCueLLM (Wu et al., 2025) introduced a lightweight plug-in that converts speech features into natural language prompts, enabling LLMs to perform multimodal emotion recognition without architectural changes.

Generation-based methods leverage LLMs to reformulate ERC as a text generation task, enabling flexible adaptation across datasets and domains. Their end-to-end nature simplifies input processing but limits fine-grained control over multimodal integration. In contrast to structured models, LLMs excel at scalability but require further refinement to model multimodal dependencies explicitly.

6 Challenges and Prospects

Based on current trends and developments in the MERC task, in this section, we outline several existing challenges and open questions, highlighting opportunities for future improvements. We structure our discussion along a logical progression: starting with foundational limitations in data collection and FAIR compliance, we then examine challenges in multimodal modeling and conclude with considerations for real-world deployment. This trajectory reflects how upstream issues in data and modeling propagate downstream, ultimately shaping the robustness, inclusivity, and applicability of MERC systems.

FAIR-related Issues Pose Challenges in MERC. The FAIR principles provide guidelines for improving the Findability, Accessibility, Interoperability, and Reusability of digital assets (Wilkinson et al., 2016). Collecting large and diverse multimodal emotion data is costly and time-

consuming; some large dialogue datasets (e.g., M^{3}ED, Zhao et al., 2022) are still monolingual and domain-bound. These limitations directly conflict with the FAIR principles. Some ERC datasets lack rich metadata or persistent identifiers, undermining findability and interoperability. Others are subject to access restrictions or copyright constraints, while many adopt inconsistent labeling schemes that hinder reusability. Consequently, researchers often have to train on small or biased samples, which undermines generalization and the reuse of models. To address these issues, future work could prioritize the development of multilingual benchmark datasets with standardized metadata and open licensing, possibly through collaborative consortiums that align with FAIR principles.

Low-Resource, Multilingual, and Multicultural Settings. As described in the previous paragraph, most state-of-the-art MERC systems are trained on English-language datasets, which limits their global applicability. Although building a large-scale, diverse MER corpus is essential, it remains an obvious challenge as it requires expert annotation of data. The limited annotated data forces researchers to rely on transfer learning (Ananthram et al., 2020), zero-shot (Qi et al., 2021), or few-shot methods (Yang et al., 2023). However, data scarcity and the high cost of emotion annotation continue to be major obstacles for MER in low-resource domains (Hussain et al., 2025). Emotions are expressed differently across languages and cultures, further compounding the challenges of MER. Variations in emotional expression and interpretation due to cultural differences can lead to inconsistencies in labeling. Most existing corpora are culture-specific, limiting their generalizability. Although researchers have acknowledged this challenge (A. and V., 2024; Ghaayathri Devi et al., 2024), MER systems aiming for global applicability must account for both linguistic diversity and culturally driven display rules. Future work could explore culture-adaptive pretraining and cross-lingual transfer learning methods that embed culturally sensitive emotion semantics across languages.

Complexities of Fusion Strategies across Modalities. Multimodal fusion techniques include early fusion, mid-level fusion, late fusion, hybrid fusion, and others (Atrey et al., 2010; Gandhi et al., 2023). A key challenge is that conversational signals, such as voice, facial expressions, and transcripts, are inherently asynchronous and occur at different time scales, making it difficult to align them at the utterance level. Emotions also depend on the context of preceding and subsequent conversation turns, so the model must capture temporal dynamics (Wang et al., 2024). Previous studies have used recurrent or self-attention layers to model sequential context (Houssein et al., 2024; Dutta et al., 2024), but long-range dependencies remain challenging to learn. How to balance and integrate contextual sentiment cue features with multimodal fusion features in decision-making, and how to determine which fusion strategies are most effective across different modalities, remain open and important research questions (A.V. et al., 2024; Ramaswamy and Palaniswamy, 2024). Recent advances in adaptive fusion strategies and dynamic attention mechanisms show promise, and future methods could explore transformer-based fusion that dynamically reweights modalities based on conversational context.

Cross-Modal Alignment, Noise Modality, Missing Modality, and Modality Conflicts. Misaligned or inconsistent features can inhibit the ability of a model to fully utilize multimodal signals, affecting its robustness and generalization (Ma et al., 2024b; Li and Tang, 2024). Noise modality, missing modality, or imbalanced modality distributions may bias simple fusion strategies (Mai et al., 2024; Zhang et al., 2024a). Even when all modalities are available, they may convey conflicting emotional signals, further complicating fusion and decision-making. Perceiving the uncertainty of different modalities for feature enhancement and resolving conflicts between modality features are important areas that need further exploration in MER research. Therefore, exploring cross-modal transfer and fusion to improve generalization in ERC has attracted the attention of more and more researchers (Fan et al., 2024; Li et al., 2024a; Feng and Fan, 2025). Some ERC methods incorporate variants of cross-modal attention (Guo et al., 2024; N. and Patil, 2020), graph-based fusion (Li et al., 2023a; Hu et al., 2021b), or mutual learning to align features during training and improve cross-domain performance (Lian et al., 2021). Future research can further investigate robust training frameworks that include modality dropout, uncertainty-aware fusion, or reinforcement learning to selectively attend to trustworthy modalities.

Effective Modality Selection. Multimodal learning refers to the integration of information from various heterogeneous sources and aims to effectively leverage data from diverse modalities (He

et al., 2024; Tsai et al., 2024). In multimodal representation learning, not all modalities contribute equally to the task. Some modalities may introduce noise and need to be removed, while others may not be essential for the task at hand but could be indispensable for other subtasks. Existing research proposes modality selection algorithms that identify the contribution of each modality (Marinov et al., 2023; Mai et al., 2024). However, selecting the most appropriate subset of modalities for a task remains one of the key challenges in multimodal learning. An emerging direction is to integrate learnable modality gates or sparsity-inducing regularization into fusion models to automatically suppress uninformative modalities during training.

Efficient Fine-tuning Approach Using Multimodal LLMs. Multimodal LLMs have brought major advances in enabling machines to learn across modalities. Some models are increasingly used in MERC, offering zero-shot or few-shot generalization across different modalities (Li et al., 2024c; Yang et al., 2024; Bo-Hao et al., 2025). The use of LLMs in MERC opens up new possibilities for capturing deeper semantic and conversational cues beyond surface-level emotion signals. However, efficiently fine-tuning these models for emotion understanding still presents challenges, especially in low-resource and culturally diverse settings. Efficient adaptation of MLLMs to capture the nuance of emotions across diverse datasets, languages, or cross-cultural settings remains an open research frontier. Promising future directions include using adapter modules or low-rank fine-tuning techniques to adapt large models to specific emotion tasks with minimal data.

MERC Application. With the growing use of interactive machine applications, MERC has become a critical research area. Applications in human-computer interaction (Ahmad et al., 2024; Moin et al., 2023), healthcare (Ayata et al., 2020; Islam et al., 2024), education (Villegas-Ch et al., 2025; Vani and Jayashree, 2025), and virtual collaboration demand robust and adaptable emotion recognition technologies that function effectively in naturalistic and dynamic environments. Yang et al. (2022) investigated MER in contexts affected by face occlusions, such as those introduced by surgical and fabric masks. Khan et al. (2024) studied contactless techniques in MER, surveying a range of nonintrusive modalities (e.g., visual cues, physiological signals). Huang (2024) developed a MER system for online learning to enable real-time monitoring and feedback on learners' emotional states. These studies highlight key directions for advancing MERC systems, particularly to make them more robust and context-aware. The ongoing research should continue to focus on real-world deployment scenarios. Future MERC systems could benefit from incremental learning techniques and user-in-the-loop feedback mechanisms that allow adaptation in dynamic, real-time environments.

Expanding the Modality Space in Emotion Recognition. Current MER systems mainly rely on vision, audio, and text, given their accessibility and prevalence in datasets. Yet human emotion is expressed through additional channels such as gaze, gestures, posture, and physiological signals (e.g., heart rate, skin conductance, brain activity) (Noroozi et al., 2021; Udahemuka et al., 2024; Wang et al., 2023; He et al., 2020; Liu et al., 2024c). These modalities remain underrepresented due to challenges in collecting synchronized, high-quality data and evolving annotation standards (Udahemuka et al., 2024; He et al., 2020; Kim and Hong, 2024). Expanding beyond the traditional three can reduce reliance on potentially misleading cues and open new application domains. For instance, biosignals and contextual cues could enhance emotion sensing in health and education, while wearable sensors and eye-trackers may enable emotion-aware experiences in VR, driver monitoring, or social robotics. In summary, gaze, physiological, and other embodied modalities are promising but underexplored in affective computing. Future work should explore how to systematically integrate these diverse modalities into large-scale benchmarks and develop models capable of robustly leveraging them in real-world scenarios.

7 Conclusion

MERC seeks to understand emotions by integrating various modalities in to dialogue of linguistic, acoustic, visual signals, and beyond. While recent progress has introduced diverse modeling strategies, significant challenges remain in data scarcity, modality alignment, and generalization across languages and cultures.

This survey provides a structured review of the MERC landscape, compares representative approaches, and highlights key open research problems. We hope it serves as a practical reference to support the future development of robust and inclusive emotion recognition systems.

Limitations

This survey offers a structured and up-to-date overview of MERC, with an emphasis on recent deep learning-based approaches. However, several limitations should be acknowledged to contextualize the scope of our work.

First, due to the focus on more advanced recent technologies, we provide only high-level summaries of representative methods, without delving into full technical details. Additionally, approaches developed prior to 2020 receive limited coverage, as our focus is primarily on recent trends that align with the rapid evolution of large-scale multimodal systems.

Second, our literature review is largely drawn from English-language publications in major conferences and repositories, including Interspeech, ICASSP, ACM, *ACL, ICML, AAAI, CVPR, COLING, and preprints on arXiv. While these venues represent core research communities in MERC, relevant contributions from other regions, languages, or domains may be underrepresented.

In the benchmark section, we highlight widely-used datasets, but do not aim for exhaustive comparison. For more in-depth benchmarking, we refer readers to complementary surveys such as Zhao et al. (2022), Sasu et al. (2025) and Gan et al. (2024).

Finally, while we identify several open challenges and underexplored directions, our discussion is not exhaustive. Rather than providing definitive answers, we aim to surface critical issues and foster further inquiry. We view these open questions as productive entry points for future research and hope this survey supports ongoing efforts to develop more advanced MERC systems.

Acknowledgments

The authors acknowledge the use of ChatGPT exclusively to refine the text for grammatical checks in the final manuscript. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011370, the National Natural Science Foundation of China (32371114), the Characteristic Innovation Projects of Guangdong Colleges and Universities (No. 2018KTSCX049), and the Guangdong Provincial Key Laboratory [No.2023B1212060076]. YL acknowledges support from the Duke Power Endowed Professorship at the School of Business, North Carolina Central University, and the IBM-HBCU Quantum Center Faculty Fellowship. ZG is supported by the National Science Foundation under ARNI (The NSF AI Institute for Artificial and Natural Intelligence) via GG019813.

References

Aruna Gladys A. and Vetriselvi V. 2024. Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning. *Applied Soft Computing*, 157:111553.

Akram Ahmad, Vaishali Singh, and Kamal Upreti. 2024. A systematic study on unimodal and multimodal human computer interface for emotion recognition. In *Computing, Internet of Things and Data Analytics*, pages 363–375, Cham. Springer Nature Switzerland.

Wei Ai, Yuntao Shou, Tao Meng, and Keqin Li. 2025a. Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4908–4921.

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025b. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11418–11426.

Amith Ananthram, Kailash Karthik Saravanakumar, Jessica Huynh, and Homayoon Beigi. 2020. Multimodal emotion detection with transfer learning. *arXiv preprint arXiv:2011.07065*.

A. Aruna Gladys and V. Vetriselvi. 2023. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, 556:126693.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379.

Geetha A.V., Mala T., Priyanka D., and Uma E. 2024. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218.

Değer Ayata, Yusuf Yaslan, and Mustafa E. Kamasak. 2020. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering*, 40(2):149–157.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

Su Bo-Hao, Shreya G. Upadhyay, and Lee Chi-Chun. 2025. Toward zero-shot speech emotion recognition using llms in the absence of target data. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, page 335–359.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4651–4660.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Jiawen Deng and Fuji Ren. 2023. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67.

Ashit Kumar Dutta, Mohan Raparthi, Mahmood Alsaadi, Mohammed Wasim Bhatt, Sarath Babu Dodda, Prashant G. C., Mukta Sandhu, and Jagdish Chandra Patni. 2024. Deep learning-based multi-head self-attention model for human epilepsy identification from eeg signal for biomedical traits. *Multimedia Tools and Applications*, 83(33):80201–80223.

Soumya Dutta and Sriram Ganapathy. 2025. Llm supervised pre-training for multimodal emotion recognition in conversations. *Preprint*, arXiv:2501.11468.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462. Association for Computing Machinery.

Yunfeng Fan, Wenchao Xu, Haohao Wang, and Song Guo. 2024. Cross-modal representation flattening for multi-modal domain generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 66773–66795. Curran Associates, Inc.

Junwei Feng and Xueyan Fan. 2025. Cross-modal context fusion and adaptive graph convolutional network for multimodal conversational emotion recognition. *arXiv preprint arXiv:2501.15063*.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3558–3565.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22):4714.

Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025a. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6748–6761, Abu Dhabi, UAE. Association for Computational Linguistics.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Yulin Wu, and Bingquan Liu. 2025b. Bemerc: Behavior-aware mllm-based framework for multimodal emotion recognition in conversation. *arXiv preprint arXiv:2503.23990*.

Chenquan Gan, Jiahao Zheng, Qingyi Zhu, Yang Cao, and Ye Zhu. 2024. A survey of dialogic emotion analysis: Developments, approaches and perspectives. *Pattern Recognition*, 156:110794.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. 2024. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 9650–9659. ACM.

K Ghaayathri Devi, Kolluru Likhitha, J Akshaya, Rfj Gokul, and G Jyothish Lal. 2024. Multi-lingual speech emotion recognition: Investigating similarities between english and german languages. In *2024 International Conference on Advances in Computing,*

Communication and Applied Informatics (ACCAI), pages 1–10.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Sreyan Ghosh, S Ramaneswaran, Utkarsh Tyagi, Harshvardhan Srivastava, Samden Lepcha, S Sakshi, and Dinesh Manocha. 2023. M-meld: A multilingual multi-party dataset for emotion recognition in conversations. *Preprint*, arXiv:2203.16799.

Ziwei Gong, Qingkai Min, and Yue Zhang. 2023. Eliciting rich positive emotions in dialogue generation. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Ziwei Gong, Muyin Yao, Xinyi Hu, Xiaoning Zhu, and Julia Hirschberg. 2024. A mapping on current classifying categories of emotions used in multimodal models for emotion recognition. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 19–28, St. Julians, Malta. Association for Computational Linguistics.

Lili Guo, Yikang Song, and Shifei Ding. 2024. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, 296:111969.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. 2024. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47):1–39.

Zhipeng He, Zina Li, Fuzhou Yang, Lei Wang, Jingcong Li, Chengju Zhou, and Jiahui Pan. 2020. Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sciences*, 10(10).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, page 1735–1780.

Essam H. Houssein, Asmaa Hammad, Nagwan Abdel Samee, Manal Abdullah Alohali, and Abdelmgeid A. Ali. 2024. Tfcnn-bigru with self-attention mechanism for automatic human emotion recognition using multi-channel eeg data. *Cluster Computing*, 27(10):14365–14385.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. 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 7042–7052, Online. Association for Computational Linguistics.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021b. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. 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 5666–5675, Online. Association for Computational Linguistics.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jian Huang, Yuanyuan Pu, Dongming Zhou, Jinde Cao, Jinjing Gu, Zhengpeng Zhao, and Dan Xu. 2024a. Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing*, 565:126992.

Ying Huang. 2024. Real-time application and effect evaluation of multimodal emotion recognition model in online learning. In *Proceedings of the 2024 10th*

International Conference on Computing and Data Engineering, ICCDE '24, page 58–63, New York, NY, USA. Association for Computing Machinery.

Zilong Huang, Man-Wai Mak, and Kong Aik Lee. 2024b. Mm-nodeformer: Node transformer multimodal fusion for emotion recognition in conversation. In *Interspeech 2024*, pages 4069–4073.

Muhammad Hussain, Caikou Chen, Sami S. Albouq, Khlood Shinan, Fatmah Alanazi, Muhammad Waseem Iqbal, and M. Usman Ashraf. 2025. Low-resource mobilebert for emotion recognition in imbalanced text datasets mitigating challenges with limited resources. *PLOS ONE*, 20(1):1–17.

Md. Milon Islam, Sheikh Nooruddin, Fakhri Karray, and Ghulam Muhammad. 2024. Enhanced multimodal emotion recognition in healthcare analytics: A deep learning based model-level fusion approach. *Biomedical Signal Processing and Control*, 94:106241.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8002–8009, Palo Alto, CA, USA. AAAI Press.

Umair Ali Khan, Qianru Xu, Yang Liu, Altti Lagstedt, Ari Alamäki, and Janne Kauttonen. 2024. Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects. *Multimedia Systems*, 30(3):115–115.

Hakpyeong Kim and Taehoon Hong. 2024. Enhancing emotion recognition using multimodal fusion of physiological, environmental, personal data. *Expert Systems with Applications*, 249:123723.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Akshi Kumar, Prakhar Dogra, and Vikrant Dabas. 2015. Emotion analysis of twitter using opinion mining. In *2015 Eighth International Conference on Contemporary Computing (IC3)*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2024a. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4):1919–1933.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023a. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427.

Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2024b. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *arXiv preprint arXiv:2407.21536*.

Jiaze Li, Hongyan Mei, Liyun Jia, and Xing Zhang. 2023b. Multimodal emotion recognition in conversation based on hypergraphs. *Electronics*, 12(22).

Songtao Li and Hao Tang. 2024. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*.

Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. 2024c. Improving context understanding in multimodal large language models via multimodal composition learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 27732–27751. PMLR.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023c. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6132–6142, New York, NY, USA. Association for Computing Machinery.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432.

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

Chenyu Liu, Xinliang Zhou, Yihao Wu, Ruizhi Yang, Zhongruo Wang, Liming Zhai, Ziyu Jia, and Yang Liu. 2024a. Graph neural networks in eeg-based emotion recognition: a survey. *arXiv preprint arXiv:2402.01138*.

Jiaxing Liu, Sheng Wu, Longbiao Wang, and Jianwu Dang. 2024b. Adaptive deep graph convolutional network for dialogical speech emotion recognition. In *Man-Machine Speech Communication*, pages 248–255, Singapore. Springer Nature Singapore.

Nian Liu, Xiao Wang, Deyu Bo, Chuan Shi, and Jian Pei. 2022. Revisiting graph contrastive learning from the perspective of graph spectrum. In *Advances in Neural Information Processing Systems*, volume 35, pages 2972–2983. Curran Associates, Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yuanyuan Liu, Lin Wei, Kejun Liu, Yibing Zhan, Zijing Chen, Zhe Chen, and Shiguang Shan. 2024c. Smile upon the face but sadness in the eyes: Emotion recognition based on facial expressions and eye behaviors. *Preprint*, arXiv:2411.05879.

Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 825–833, New York, NY, USA. Association for Computing Machinery.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024a. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Wenxuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. 2024b. Learning modality knowledge alignment for cross-modality transfer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33777–33793. PMLR.

Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. 2024. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, 26:3018–3033.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 6818–6825.

Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. DialogueTRM: Exploring multimodal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zdravko Marinov, Alina Roitberg, David Schneider, and Rainer Stiefelhagen. 2023. Modselect: Automatic modality selection for synthetic-to-real domain generalization. In *Computer Vision – ECCV 2022 Workshops*, pages 326–346, Cham. Springer Nature Switzerland.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *SciPy*.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Anam Moin, Farhan Aadil, Zeeshan Ali, and Dongwann Kang. 2023. Emotion recognition framework using multiple modalities for an effective human–computer interaction. *The Journal of Supercomputing*, 79(8):9320–9349.

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

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, Singapore. Association for Computational Linguistics.

Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2021. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2):505–523.

Sancheng Peng, Lihong Cao, Yongmei Zhou, Zhouhao Ouyang, Aimin Yang, Xinguang Li, Weijia Jia, and Shui Yu. 2022. A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, 8(5):745–762.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. 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, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2021. Zero-shot video emotion recognition via multimodal protagonist-aware transformer network. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1074–1083, New York, NY, USA. Association for Computing Machinery.

Manju Priya Arthanarisamy Ramaswamy and Suja Palaniswamy. 2024. Multimodal emotion recognition: A comprehensive review, trends, and challenges. *WIREs Data Mining and Knowledge Discovery*, 14(6):e1563.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. 2025. Akan cinematic emotions (ACE): A multimodal multi-party dataset for emotion recognition in movie dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9820–9831, Vienna, Austria. Association for Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Björn W. Schuller, Michel François Valstar, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012: the continuous audio/visual emotion challenge - an introduction. In *International Conference on Multimodal Interaction, ICMI '12, Santa Monica, CA, USA, October 22-26, 2012*, pages 361–362. ACM.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 493–502, New York, NY, USA. Association for Computing Machinery.

Tao Shi and Shao-Lun Huang. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Yuntao Shou, Wei Ai, Jiayi Du, Tao Meng, Haiyan Liu, and Nan Yin. 2024. Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. *arXiv preprint arXiv:2407.00119*.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. 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, Abu Dhabi, UAE. Association for Computational Linguistics.

Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan. 2024. Low-rank approximation for sparse attention in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13763–13773.

Yanyuan Su, Cuijuan Han, Yaming Zhang, and Haiou Liu. 2024. Multimodal sentiment recognition driven by feature fusion strategy for social media comments on sudden natural disasters. *Available at SSRN 5011893*.

Hiroki Tanioka, Tetsushi Ueta, and Masahiko Sano. 2024. Toward a dialogue system using a large language model to recognize user emotions with a camera. *arXiv preprint arXiv:2408.07982*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks . In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, Los Alamitos, CA, USA. IEEE Computer Society.

Yun-Da Tsai, Ting-Yu Yen, Pei-Fu Guo, Zhe-Yan Li, and Shou-De Lin. 2024. Text-centric alignment for multi-modality learning. *arXiv preprint arXiv:2402.08086*.

Gustave Udahemuka, Karim Djouani, and Anish M. Kurien. 2024. Multimodal emotion recognition using visual, vocal and physiological signals: A review. *Applied Sciences*, 14(17).

Cuong Tran Van, Thanh V. T. Tran, Van Nguyen, and Truong Son Hy. 2025. Effective context modeling framework for emotion recognition in conversations. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

RK Vani and P Jayashree. 2025. Multimodal emotion recognition system for e-learning platform. *Education and Information Technologies*, pages 1–32.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

William Villegas-Ch, Rommel Gutierrez, and Aracely Mera-Navarrete. 2025. Multimodal emotional detection system for virtual educational environments: Integration into microsoft teams to improve student engagement. *IEEE Access*, 13:42910–42933.

Ling Wang, Jiayu Hao, and Tie Hua Zhou. 2023. Ecg multi-emotion recognition based on heart rate variability signal features mining. *Sensors*, 23(20).

Peng Wang, Lesya Ganushchak, Camille Welie, and Roel van Steensel. 2024. The dynamic nature of emotions in language learning context: Theory, method, and analysis. *Educational Psychology Review*, 36(4):105.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Mark D. Wilkinson, Michel Dumontier, I. J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, and Erik Schultes. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

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

Yunhe Xie, Kailai Yang, Chengjie Sun, Bingquan Liu, and Zhenzhou Ji. 2021. Knowledge-interactive network with sentiment polarity intensity-aware multitask learning for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2879–2889, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. 2021. Hierarchical self-supervised augmented knowledge distillation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1217–1223. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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

Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Soujanya Poria. 2023. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6045–6053, New York, NY, USA. Association for Computing Machinery.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2025. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25642–25650.

Ziqing Yang, Katherine Nayan, Zehao Fan, and Houwei Cao. 2022. Multimodal emotion recognition with surgical and fabric masks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4678–4682.

Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. 2023. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. In *Advances in Neural Information Processing Systems*, volume 36, pages 69638–69660. Curran Associates, Inc.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *CoRR*, abs/1708.04299.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024a. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Tao Zhang and Zhenhua Tan. 2024. Survey of deep emotion recognition in dynamic data using facial, speech and textual cues. *Multimedia Tools and Applications*, 83(25):66223–66262.

Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 13099–13110, Toronto, Canada. Association for Computational Linguistics.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, 192:107901.

Zhengxuan Zhang, Jianying Chen, Xuejie Liu, Weixing Mai, and Qianhua Cai. 2024b. 'what' and 'where' both matter: dual cross-modal graph convolutional networks for multimodal named entity recognition. *International Journal of Machine Learning and Cybernetics*, 15(6):2399–2409.

Zhengxuan Zhang, Weixing Mai, Haoliang Xiong, Chuhan Wu, and Yun Xue. 2023. A token-wise graph-based framework for multimodal named entity recognition. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2153–2158.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.

Xiaofei Zhu, Jiawei Cheng, Zhou Yang, Zhuo Chen, Qingyang Wang, and Jianfeng Yao. 2024. Cmath: Cross-modality augmented transformer with hierarchical variational distillation for multimodal emotion recognition in conversation. *arXiv preprint arXiv:2411.10060*.

Shihao Zou, Xianying Huang, and Xudong Shen. 2023. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5994–6003, New York, NY, USA. Association for Computing Machinery.

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

A Datasets Details

IEMOCAP. The IEMOCAP dataset (Busso et al., 2008) consists of dyadic conversation videos from ten speakers, comprising 151 dialogues and 7,433 utterances. Sessions 1 to 4 are used as the training set, while the last session is held out as the test set. Each utterance is annotated with one of six emotion labels: happy, sad, neutral, angry, excited, and frustrated.

MELD. The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019a) is an extension of the EmotionLines corpus (Hsu et al., 2018), constructed from the TV series Friends. It contains 1,433 multi-party conversations and 13,708 utterances. Each utterance is annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise. Unlike dyadic datasets, MELD captures the complexity of multi-speaker interactions, making it well-suited for studying emotion recognition in multi-party conversational settings.

CMU-MOSEI. The CMU-MOSEI dataset (Bagher Zadeh et al., 2018) consists of 23,453 sentence-level video segments from over 1,000 speakers covering 250 topics, collected from YouTube monologue videos. Each segment is annotated for sentiment on a 7-point Likert scale ([-3: highly negative, -2: negative, -1: weakly negative, 0: neutral, +1: weakly positive, +2: positive, +3: highly positive]) and for the intensity of six Ekman (Ekman et al., 1980) emotions: happiness, sadness, anger, fear, disgust, and surprise. The dataset provides aligned language, visual, and acoustic modalities, making it a large-scale benchmark for multimodal sentiment and emotion recognition.

M^{3}ED. The M^{3}ED dataset (Zhao et al., 2022) consists of 990 dyadic dialogues and 24,449 utterances collected from 56 Chinese TV series. Each utterance is annotated with one or more of seven emotion labels: happy, surprise, sad, disgust, anger, fear, and neutral. The dataset covers text, audio, and visual modalities and features blended emotions and speaker metadata, making it the first large-scale multimodal emotional dialogue corpus in Chinese.

ACE. The ACE dataset (Sasu et al., 2025) contains 385 emotion-labeled dialogues and 6,162 utterances in the Akan language, collected from 21 movie sources. It includes multimodal information across text, audio, and visual modalities, and is annotated with one of seven emotion categories: neutral, sadness, anger, fear, surprise, disgust, and happiness. Word-level prosodic prominence is also provided, making it the first such dataset for an African language. The dataset is gender-balanced, featuring 308 speakers, and is split into training, validation, and test sets in a 7:1.5:1.5 ratio.

MEmoR. The MEmoR dataset (Shen et al., 2020) consists of 5,502 video clips and 8,536 person-level samples extracted from the sitcom The Big Bang Theory, annotated with 14 fine-grained emotions. Unlike most datasets focusing solely on speakers, MEmoR includes both speakers and non-speakers, even when modalities are partially or completely missing. Each sample comprises a target person, an emotion moment, and multimodal inputs (text, audio, visual, and personality features), making it a challenging benchmark for multimodal emotion reasoning beyond direct recognition.

AVEC. The AVEC dataset (Schuller et al., 2012), derived from the SEMAINE corpus (McKeown et al., 2012), features human-agent interactions annotated with four continuous affective dimensions: valence, arousal, expectancy, and power. While the original labels are provided at 0.2-second intervals, we aggregate them over each utterance to obtain utterance-level annotations for emotion analysis.

M-MELD. M-MELD (Ghosh et al., 2023) is a multilingual extension of the MELD dataset, created to support emotion recognition in conversations across different languages. While MELD is an English-only multimodal dataset, M-MELD includes human-translated versions of the original utterances in four additional languages: French, Spanish, Greek, and Polish. This multilingual corpus retains the original multimodal structure and emotion labels, enabling research in cross-lingual and multimodal emotion analysis. By balancing high-resource and low-resource languages, M-MELD offers a valuable benchmark for developing and evaluating multilingual ERC models.

EmoryNLP. EmoryNLP (Zahiri and Choi, 2017) is a textual emotion-labeled dataset derived from the Friends TV series, containing over 12,000 utterances across multi-party dialogues. Each utterance is annotated with one of seven emotions: neutral, joyful, peaceful, powerful, mad, sad, or scared, offering a fine-grained resource for emotion recognition in conversational settings.