MATCH: Modality-Calibrated Hypergraph Fusion Network for Conversational Emotion Recognition

Jiandong Shi[1], **Ming Li**[2,3,*], **Lu Bai**[4], **Feilong Cao**[5], **Ke Lu**[6,7], **Jiye Liang**[8]
[1]School of Computer Science and Technology, Zhejiang Normal University
[2]Zhejiang Key Laboratory of Intelligent Education Technology and Application Zhejiang Normal University
[3]Zhejiang Institute of Optoelectronics
[4]School of Artificial Intelligence, Beijing Normal University
[5]School of Mathematical Sciences, Zhejiang Normal University
[6]School of Engineering Science, University of Chinese Academy of Sciences
[7]Peng Cheng Laboratory
[8]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, the School of Computer and Information Technology, Shanxi University
{shijiandong, mingli, caofeilong88}@zjnu.edu.cn, bailu@bnu.edu.cn, luk@ucas.ac.cn, ljy@sxu.edu.cn

Abstract

Multimodal emotion recognition aims to identify emotions by integrating multimodal features derived from spoken utterances. However, existing work often neglects the calibration of conversational entities, focusing mainly on extracting potential intra- or cross-modal information. This leads to the underutilization of utterance information that is essential for accurately characterizing emotion. Additionally, the lack of effective modeling of conversational patterns limits the ability to capture emotional pathways across contexts, modalities and speakers, impacting the overall emotional understanding. In this study, we propose the modality-calibrated hypergraph fusion network (**MATCH**), which leverages multimodal fusion and hypergraph learning techniques to address these challenges. In particular, we introduce an entity calibration strategy that refines the representations of conversational entities both at the modality and context levels, allowing for deeper insights into emotion-related cues. Furthermore, we present an emotion-aligned hypergraph fusion method that incorporates a line graph to explore conversational patterns, facilitating flexible knowledge transfer across modalities through hyperedge-level and graph-level alignments. Experiments demonstrate that **MATCH** outperforms state-of-the-art approaches on two benchmark datasets.

1 Introduction

Emotion recognition in conversations (ERC) aims to detect emotional states from conversational signals, providing crucial emotional cues for downstream tasks. ERC has attracted significant attention in fields such as social recommendation [Zhang *et al.*, 2024], fake news detection [Mittal *et al.*, 2020], and dialogue systems [Bertero *et al.*, 2016]. Early ERC approaches primarily focused on text-based inputs, utilizing techniques like recurrent neural networks (RNNs) [Majumder *et al.*, 2019], Transformers [Lian *et al.*, 2021], and graph neural networks (GNNs) [Ghosal *et al.*, 2019] to extract emotional features from text, achieving some success. Recently, with the increased availability of multimodal data, research has shifted towards multimodal emotion recognition in conversations (MERC), aiming to enhance emotional understanding by integrating and analyzing information across multiple modalities.

Existing MERC approaches can be broadly classified into aggregation-based methods and graph-based methods. Aggregation-based methods combine modality information using techniques such as concatenation, attention mechanisms, and tensor fusion to perform emotion prediction [Zadeh *et al.*, 2017; Zadeh *et al.*, 2018]. In contrast, graph-based methods capture modality interactions and contextual dependencies in conversations through node-level propagation and diverse edge designs [Ghosal *et al.*, 2019; Hu *et al.*, 2022; Shi *et al.*, 2025]. While effective, graph-based methods struggle to model complex multivariate dependencies between utterances, leading to the development of hypergraph-based approaches. For instance, M^{3}Net [Chen *et al.*, 2023] captures the multivariate and multi-frequency characteristics of multimodal features by integrating hypergraph and frequency domain decomposition on graph, while HAUCL [Yi *et al.*, 2024] uses a hypergraph autoencoder to learn adaptive hyperedge connectivity patterns that are relevant for emotion prediction.

Despite these advancements, there exist two key challenges: (i) Insufficient calibration of modality features: Raw modality features often encompass intricate and multifaceted

*Corresponding Author (mingli@zjnu.edu.cn)

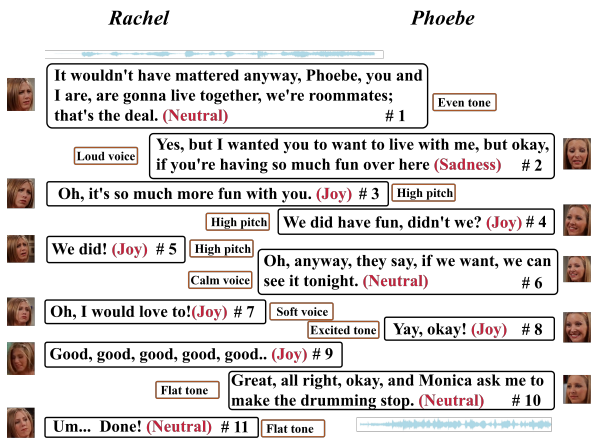

Figure 1: Illustrations of conversational patterns in multimodal scenarios, highlighting emotional pathways across modalities, contexts, and speaker information. The golden labels are highlighted in red.

information [Li *et al.*, 2023a]; however, existing approaches either overlook deeper exploration of these features or simply categorize them into intra-modal and inter-modal information. This not only limits the utilization of modal features but may also result in the misinterpretation of emotional cues during prediction due to imprecise delineation of these feature types. (ii) Inadequate learning of conversational patterns: Real-world conversations exhibit inherent structural patterns across utterances. For instance, a person expressing happiness may simultaneously smile and use a high-pitched voice, which are correlated features that may appear in utterances across different speakers, influencing emotional states. Such conversational patterns reflect the dynamic pathways of emotion shifts across context and modalities, which are critical for MERC tasks. However, existing approaches struggle to effectively capture and represent these complex pathways, resulting in superficial emotional understanding.

As illustrated in Figure 1, in utterance #4, **Phoebe** says *"We did have fun, didn't we?"*, accompanied by a smile and a high-pitched tone, which influences **Rachel**, who also smiles and uses a high-pitched tone. This expression of joy is difficult to identify from the text alone, as there is no direct contextual connection between the two. Similarly, in utterances #10 and #11, **Phoebe**'s intonation and expression affect **Rachel**. Thus, the emotional state of utterance #11 shifts to neutral, rather than maintaining the previously expressed joy. These nuances of emotion across utterance, modality, and speaker require a deeper exploration and calibration of modality features to prevent interference from erroneous contextual and modality information. More importantly, they highlight emotional pathways between utterances, enhancing a deeper understanding while minimizing the impact of unnecessary contextual or modality information, emphasizing the need to explore conversational patterns.

To address these challenges, we propose a modality-calibrated hypergraph fusion network, named **MATCH**. **MATCH** comprises two key components: conversational entity calibration and emotion-aligned hypergraph fusion. The entity calibration strategy focuses on the critical entities in ERC tasks, i.e., utterance and speaker, and performs a fine-grained calibration at the modality and context levels. This strategy yields a more refined multimodal representation compared to simply distinguishing between intra- and inter-modal information. Subsequently, the calibrated features are used to construct a hypergraph that captures high-order semantic relationships. Moreover, we construct a line graph to extract conversational patterns that are challenging to be represented by hypergraphs. Together, these components enable **MATCH** to perceive both surface-level semantic and deeper emotional pathways, facilitating the generation of comprehensive emotional representations. In summary, our contributions are as follows:

- We propose **MATCH**, a hypergraph-based MERC model that delivers comprehensive emotional understanding by conversational entity calibration and emotion-aligned hypergraph fusion.
- We design a fine-grained conversational entity calibration strategy that enhances the utilization of multimodal features by calibrating utterance and speaker knowledge at both the modality and context levels.
- We propose emotion-aligned learning, which maximizes the role of hyperedges in MERC by hyperedge- and graph-level alignment. This facilitates the learning of surface semantics and deep emotional pathways, enhancing the emotional understanding.

2 Related Work

2.1 Multimodal Fusion

Multimodal fusion aims to produce a more comprehensive representation by integrating multimodal information through early fusion, decision fusion, and hybrid fusion strategies [Zhao *et al.*, 2024]. Early fusion combines modalities into a joint representation [Mai *et al.*, 2020], while decision fusion aggregates predictions from individual modalities using weighted summation or expert voting. Hybrid fusion blends the advantages of both approaches, offering greater flexibility [Duan *et al.*, 2024; Tellamekala *et al.*, 2023; Li *et al.*, 2025d]. Recent advances in graph deep learning have enhanced modality interaction capture in MERC. However, a common issue in both graph-based and non-graph-based methods is the lack of modality calibration, as simply extracting inter- or intra-modal information is often insufficient for producing appropriate results [Joshi *et al.*, 2022; Li *et al.*, 2023b; Li *et al.*, 2023a]. This challenge aligns with real-world conversations, where emotional states conveyed by an utterance are not solely expressed through modality information.

2.2 Hypergraphs in Emotion Recognition

Hypergraphs provide a powerful framework for modeling high-order interactions among multiple entities, going beyond the pairwise correlations captured by traditional graphs [Millán *et al.*, 2025]. This flexibility makes them especially effective for representing complex relationships in a wide range of real-world datasets [Feng *et al.*, 2019; Ju *et al.*, 2024; Li *et al.*, 2025c; Li *et al.*, 2025b; Li *et al.*, 2024;

Li *et al.*, 2025a]. Recent studies have leveraged this capability to MERC tasks, aiming to capture high-order information through a multi-node connectivity paradigm [Chen *et al.*, 2023; Yi *et al.*, 2024]. However, an important challenge is that hyperedges, as part of the hypergraph, reflect its inherently valuable attributes [Wang *et al.*, 2024; Chen *et al.*, 2024]. Existing research has not explored how to leverage hyperedge information to enrich emotion representations [Lu *et al.*, 2024], limiting the potential of hypergraphs for MERC tasks. Meanwhile, the issue of information imbalance between modalities persists, causing some hyperedges to propagate weak semantic knowledge, resulting in significant bias in emotional prediction.

3 Methodology

In this section, we provide preliminaries and a detailed introduction to each component of the proposed **MATCH**, as depicted in Figure 2.

3.1 Task Definition

Let $\mathcal{C} = [c_1, c_2, \cdots, c_{\mathcal{N}}]$ denote a set of $\mathcal{N}$ conversations and $U = [u_1, u_2, \cdots, u_{\mathcal{M}}]$ represent a set of $\mathcal{M}$ utterances, respectively. Each conversation c_i consists of $\mathcal{M}$ utterances and involves $\mathcal{Q} \geq 2$ speakers. An utterance u_i, is represented as a triplet $u_i = \{u_i^a, u_i^v, u_i^t\}$, where $u_i^a \in \mathbb{R}^{d_a}$, $u_i^v \in \mathbb{R}^{d_v}$ and $u_i^t \in \mathbb{R}^{d_t}$ represent the acoustic, visual, and textual features of u_i, respectively. MERC aims to predict the emotional label $\hat{y}_i$ for each utterance u_k based on its corresponding triplet representation.

3.2 Preliminaries

Definition 1. ***Graph*** *follows the paradigm of pairwise node connections. A traditional graph (or in short, a graph) can be defined as: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ and $\mathcal{E}$ denotes a set of vertices and edges, respectively. $X_g \in \mathbb{R}^{|\mathcal{V}| \times d_h}$ is the feature matrix, d_h is the dimension of features, the adjacency matrix of $\mathcal{G}$ is $\Lambda_g \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$.*

Definition 2. ***Hypergraph*** *extends beyond the pairwise connection paradigm, offering a high-order representation structure. A hypergraph can be defined as: $\mathcal{G}_h = (\mathcal{V}, \mathcal{E}_h, \mathcal{W})$ where $\mathcal{V}$ is a vertices set initialized with feature $X_h \in \mathbb{R}^{|\mathcal{V}| \times d_h}$ and $\mathcal{E}_h$ is a hyperedges set which contains multiple vertices $\{v_1, \cdots, v_n\}$. $\Lambda \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}_h|}$ is the corresponding incidence matrix.*

Definition 3. ***Line Graph*** *is derived from hypergraph $\mathcal{G}_h$ to capture the structured information within $\mathcal{G}_h$. A line graph can be defined as: $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$, where each vertex $v_i \in \mathcal{V}_l$ is a vertex-hyperedge pair $\{(v, e) \mid v \in e, v \in \mathcal{V}_h, e \in \mathcal{E}_h\}$ from $\mathcal{G}_h$. Edge set $\mathcal{E}_l$ and adjacency matrix $\Lambda_l \in \{0, 1\}$ is defined by the relation with $\Lambda_l(v_l, v_l) = 1$ if either $v = v'$ or $e = e'$ for $v_l = (v, e), v_l = (v', e') \in \mathcal{V}_l$.*

3.3 Utterance Encoding

For visual and acoustic features, we use two separate fully connected layers to obtain their respective representations.

$$c_i^\zeta = \text{FC}^\zeta(u_i^\zeta; \theta_{\text{FC}}^\zeta), \ \ \zeta \in \{a, v\}, \tag{1}$$

where c_i^ζ denotes the representation for utterance u_i^ζ. θ_{FC}^ζ are learnable parameters. For textual features, a bidirectional GRU is employed to enhance contextual coherence and obtain the corresponding representations:

$$c_i^t, h_i^t = \overleftrightarrow{GRU}^t(u_i^t, h_k^t), \ \ k<i, \tag{2}$$

where h_i^t is hidden state of the i-th utterance.

3.4 Conversational Entity Calibration

For MERC, entity information in a conversation carries varying levels of emotional cues. Unlike previous work that defines all possible relationships in a dialogue as entities (e.g., between speakers or within the same speaker), we categorize entity information into two types: (i) utterance information and (ii) speaker information. After utterance encoding, we obtain the contextual information c_i^ξ corresponding to each utterance. For speaker information, we also utilize a bidirectional GRU to extract it.

$$s_i^\xi, \hat{h}_i^\xi = \overleftrightarrow{GRU^\xi}(u_i^\xi, \hat{h}_k^\xi), \ \ \xi \in \{a, v, t\}, \tag{3}$$

where s_i^ξ denotes the speaker information corresponding to u_i^ξ and $\hat{h}_i^\xi$ is the hidden state.

Previous work has focused solely on decoupling utterance information [Li *et al.*, 2023a], which proves insufficient in scenarios where the emotional states of speakers vary dynamically. To address this, we calibrate both two conversational entities, thereby constructing a more precise emotional learning space. Given the inherent differences in information density across modalities, we first leverage text to augment the acoustic and visual features. Take speaker representations s_i^ξ for instance:

$$\hat{s}_i^v = \sum_{k=1}^{\mathcal{M}} \frac{\exp(\text{sim}(s_i^t, s_k^v)/\tau_1)}{\sum_{j=1}^{\mathcal{M}} \exp(\text{sim}(s_i^t, s_j^v)/\tau_1)} * s_k^v, \tag{4}$$

where $\hat{s}_i^v$ denotes the enhanced visual representations. sim(★,★) is the cosine similarity function and τ_1 is a temperature parameter. The enhanced acoustic representations $\hat{s}_i^a$ can be captured in the same manner.

The speaker information often overlaps across both contextual and modality levels, which can obscure clear emotional cues. Therefore, we calibrate speaker information at both levels to ensure a more accurate and coherent emotional representation. Specifically, we enhance recognition performance by adjusting the distance between semantically similar representations in the semantic space through contrastive learning. First, we add Gaussian noise as an effective way to mitigate the inevitable loss of information noted in [Wang *et al.*, 2023] and re-normalize the features when projecting modality data into the semantic space:

$$\tilde{s}_i^\xi = \text{Norm}(\hat{s}_i^\xi + \theta^\xi), \tag{5}$$

where noise $\theta^\xi, \xi \in \{a, v, t\}$ is sampled from zero-mean gaussian distribution. Then, the modality- and contextual-level calibration losses can be defined as:

$$\mathcal{L}_{cal}^m = -\sum_{i=1}^{\mathcal{M}} \sum_{k \in i^{m+}}^{\mathcal{M}} \log \frac{\exp(\text{sim}(\tilde{s}_i^{m,\xi}, \tilde{s}_k^{m,\xi})/\tau_m)}{\sum_{j \neq i}^{3\mathcal{M}} \exp(\text{sim}(\tilde{s}_i^{m,\xi}, \tilde{s}_j^{m,\xi})/\tau_m)}, \tag{6}$$

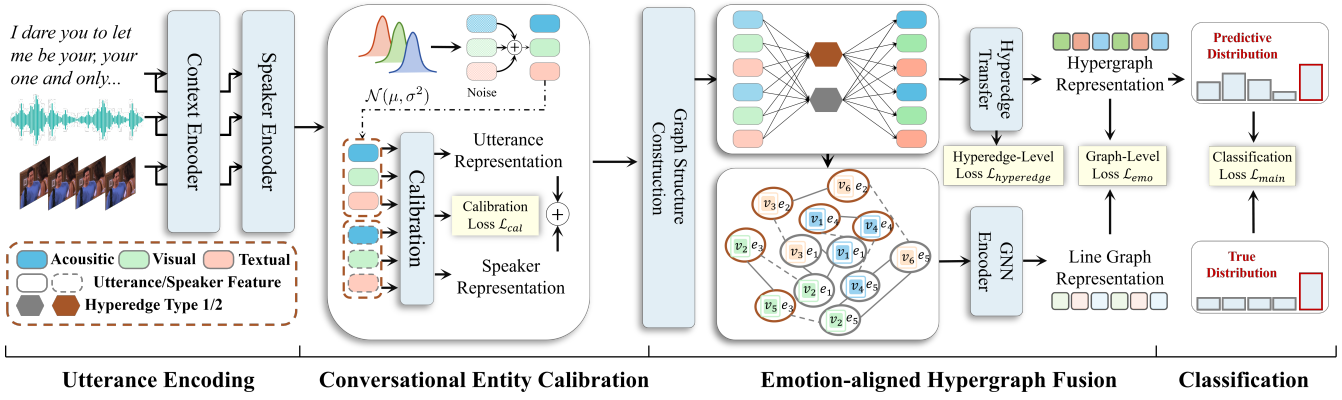

Figure 2: Schematic of the proposed MATCH framework.

$$\mathcal{L}_{cal}^{c} = -\sum_{i=1}^{\mathcal{M}} \sum_{k \in i^{c+}}^{\mathcal{M}} \log \frac{\exp(\text{sim}(\tilde{s}_i^{c,\xi}, \tilde{s}_k^{c,\xi})/\tau_c)}{\sum_{j \neq i}^{3\mathcal{M}} \exp(\text{sim}(\tilde{s}_i^{c,\xi}, \tilde{s}_j^{c,\xi})/\tau_c)}, \quad (7)$$

where $\tilde{s}_i^{m,\xi} = \mathcal{T}_m(\tilde{s}_i^{\xi})$ and $\tilde{s}_i^{c,\xi} = \mathcal{T}_c(\tilde{s}_i^{\xi})$. $\mathcal{T}(\bigstar)$ denotes the projector, τ_m, τ_c are temperature parameters. i^{m+}, i^{c+} denote the positive modality- and context-level list for i-th sample. In this manner, the modality and contextual cues are distributed across different locations in the representation space, making it easier to calibrate the deeper semantics conveyed by each speaker. Combining Equation. (4)-(7), we can obtain the calibrated representations as:

$$\overline{s}_i^{\xi} = s_i^{\xi} \oplus \tilde{s}_i^{m,\xi} \oplus \tilde{s}_i^{c,\xi}. \quad (8)$$

We apply the same process to the utterance information c_i^{ξ} to obtain the calibrated utterance representations $\overline{c}_i^{\xi}$. The final calibrated representation for the i-th utterance is:

$$x_i^{\xi} = \overline{c}_i^{\xi} + \eta * \overline{s}_i^{\xi}, \ \ x_i^{\xi} \in \mathbb{R}^{d_h}, \quad (9)$$

where η is a hyperparameter to control the weight of calibrated speaker information. The overall calibration loss is:

$$\mathcal{L}_{cal} = \sum_{i \in \{u,s\}} \sum_{j \in \{m,c\}} \mathcal{L}_{cal}^{i,j}. \quad (10)$$

3.5 Emotional-aligned Hypergraph Fusion

Graph structures are effective in ERC tasks due to their ability to model non-Euclidean data. However, traditional graph neural networks have an inherent limitation, i.e., message propagation occurs point-to-point, which leads to a loss of context during the learning process. In contrast, hypergraphs overcome this by forming hyperedges that connect multiple nodes, enabling the capture of multivariate conversational relationships that are difficult to model with traditional graphs.

Existing hypergraph-based approaches face a significant challenge: the underutilization of hyperedge representations, leading to shallow and incomplete emotion understanding. These approaches primarily focus on node-level representations, while node information is eventually aggregated into hyperedge representations and propagated back to the nodes. The weak semantic nature of the hyperedges dilutes their impact on individual nodes, causing semantic imbalance. For instance, nodes connected to textual modality hyperedges typically exhibit stronger semantics than those linked to visual or acoustic hyperedges.

To this end, we introduce the line expansion [Yang *et al.*, 2022] to unleash the power of hypergraph for MERC. First, we categorize emotional features into two main types: (i) semantic information learned by hypergraph, and (ii) conversational pattern learned by line graph. These two types of information complement each other, revealing emotional pathways between utterances and enabling the generation of a more comprehensive understanding of emotion. Additionally, we introduce emotion-aligned learning to ensure consistent understanding across these two perspectives.

Hypergraph Semantic Learning. We begin by constructing conversational hypergraph $\mathcal{G}_h = (\mathcal{V}, \mathcal{E}_h)$ to capture high-order semantic nuances across utterances. Following [Chen *et al.*, 2023], we define two types of hyperedges $(e_1, e_2) \in \mathcal{E}_h$: one connecting nodes within the same context and the other connecting nodes within the same modality. Each item in triplets of utterances $\{u_i^a, u_i^v, u_i^t\}$ is treated as nodes $v \in \mathcal{V}, |\mathcal{V}| = 3 \times \mathcal{M}$, which are initialized with calibrated representations x_i^{ξ}. Subsequently, we have the corresponding hypergraph incidence matrix $\Lambda \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}|}$. Based on this, we apply a hypergraph convolutional network to capture the fundamental semantic representations shared between utterances. Formally:

$$X^{(l+1)} = \sigma(D_v^{-1/2} \Lambda W D_e^{-1} \Lambda^{\top} D_v^{-1/2} X^{(l)} \Theta^{(l)}), \quad (11)$$

where D_v and D_e denote the node degree matrix and hyperedge degree matrix, respectively. σ represents a nonlinear activation function. W is a learnable matrix. Considering that the low density of semantics in non-textual modalities may be further amplified during the propagation, we adopt an emotion-aligned learning strategy to alleviate this issue, which will be presented in the following sections.

Conversational Pattern Learning. Although multimodal representations learned through hypergraphs can capture emotion states to some extent, meaningful emotion consensus can also emerge between unpaired utterances in complex dialogues through cross-modal information, i.e., the pathways

of emotion transfer within the context. These emotion pathways are reflected in the information flow across hyperedges, while hypergraph is more sensitive to node-level details. To address this, we introduce line graphs to further learn the conversational pattern. First, we construct the line graph adjacency matrix $\hat{\Lambda} \in \mathbb{R}^{|\mathcal{R}|\times|\mathcal{R}|}$ based on $\Lambda \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$, where $\mathcal{R} = 6 \times \mathcal{M}$. Subsequently, we transform the hypergraph node embedding $X \in \mathbb{R}^{|\mathcal{V}|\times d_h}$ into the line graph feature vector through a node projector $\mathcal{P}_v$ to explore the conversational pattern:

$$\hat{X} = \mathcal{P}_v X \in \mathbb{R}^{|\mathcal{R}|\times d_h}. \tag{12}$$

Let $\hat{X}$ be the line graph feature at the first layer, that is, $\hat{X}^{(0)} = \hat{X}$, the following convolution on line graph is:

$$\hat{X}^{(k+1)} = \sigma(\tilde{D}^{-1/2}\tilde{\Lambda}D^{-1/2}\hat{X}^{(k)}W_{le}^{(k)}), \tag{13}$$

where $\tilde{\Lambda} = \hat{\Lambda} + 2\mathrm{I}$, W_{le} is a learnable matrix.

Let $\hat{X}^{(K)}$ denote the node representations obtained from the line graph learning module. To preserve semantic consistency, we apply the back-projector $\mathcal{P}_v^{'}$, which re-projects the line node representations into the hypergraph space. This reprojection is performed based on the reciprocal of the edge degree, formally expressed as:

$$\overline{X}^{(K)} = \mathcal{P}_v^{'}\hat{X}^{(K)} \in \mathbb{R}^{|\mathcal{V}|\times d_h}. \tag{14}$$

In this way, the structural information embedded in the hypergraph is re-expressed as vertices in the line graph through the convolution and back-projection processes. This enables the model to capture and understand the deeper, hidden conversational patterns.

Emotional-aligned Learning. Hypergraphs and line graphs provide complementary perspectives for emotion understanding. However, weak semantic hyperedges reduce their ability to effectively convey emotion. To address this, we propose emotion-aligned learning that enhances the effectiveness of semantic transfer by optimizing the distribution discrepancy through hyperedge- and graph-level alignment. Hyperedge-level alignment optimizes the gap between weak and strong semantic hyperedges, enhancing the representational capacity of acoustic and visual features through the incorporation of textual features. Based on Equation. (11), the hyperedge representations can be defined as follows:

$$\mathcal{B} = \Lambda^{\top} D_v^{-1/2} X \Theta \in \mathbb{R}^{|\mathcal{E}|\times d_h}, \tag{15}$$

where $\mathcal{B}$ denotes the hyperedge representations, let $\mathcal{B}^{\xi}$ denotes the representations of three modalities. The corresponding distributions are:

$$\mathcal{D}^{\xi} = \mathrm{softmax}(\mathcal{B}^{\xi}/\tau_{emo}), \tag{16}$$

where τ_{emo} is a temperature parameter.

We introduce the Kullback-Leibler (KL) divergence loss to minimize the distribution gap, formally:

$$\mathcal{L}_{hyperedge} = \sum_{\zeta\in\{a,v\}} \mathcal{D}^{t} \log\left(\frac{\mathcal{D}^{t}}{\mathcal{D}^{\zeta}}\right). \tag{17}$$

In this manner, the hyperedges $\mathcal{B}^{\zeta}$ enrich their representations by iterating through the strong semantic representation, preventing the propagation of invalid information.

Note that, both line graph and hypergraph provide meaningful emotional representations, and the graph-level alignment aims to enhance knowledge transfer between line graph and hypergraph, formally:

$$\begin{aligned}\mathcal{L}_{emo} = &\frac{1}{2\mathcal{M}}\sum_{j=1}^{\mathcal{M}}\left(\overline{\mathscr{X}_j}^{(K)}\log\left(\frac{\overline{\mathscr{X}_j}^{(K)}}{\mathscr{X}_j^{(L)}}\right)\right.\\ &\left.+\mathscr{X}_j^{(L)}\log\left(\frac{\mathscr{X}_j^{(L)}}{\overline{\mathscr{X}_j}^{(K)}}\right)\right),\end{aligned} \tag{18}$$

where $\mathscr{X}$ and $\overline{\mathscr{X}}$ are the distributions derived from Equation (16) corresponding to X and $\overline{X}$.

3.6 Training Objective

Finally, we apply a fully connected layer to obtain the prediction labels $\hat{y}_i$. Formally:

$$\hat{y}_i = \mathrm{softmax}(W_{cls}(x_i^{(L)} \oplus \overline{x}_i^{(K)}) + b_{cls}). \tag{19}$$

The main ERC task loss can be defined as:

$$\mathcal{L}_{main} = -\frac{1}{\mathcal{N}\times\mathcal{M}}\sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{\mathcal{M}} y_{i,j}\log(\hat{y}_{i,j}) + \eta_2\|\Theta\|_2, \tag{20}$$

where η_2 is a hyperparameter. The overall loss is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_{main} + \gamma_1 * \mathcal{L}_{cal} + \gamma_2 * \mathcal{L}_{hyperedge} + \gamma_3 * \mathcal{L}_{emo}, \tag{21}$$

where $\gamma_1, \gamma_2, \gamma_3$ are hyperparameters. Their sensitivity is studied in Section 4.6.

4 Experiments

4.1 Datasets

We evaluate the **MATCH** on two benchmark datasets, IEMOCAP and MELD. The detailed statistics are shown in Table 1.

IEMOCAP [Busso *et al.*, 2008] contains video data from dyadic conversations with ten speakers, with utterances classified into six emotion categories: *Happy, Sad, Neutral, Angry, Excited, Frustrated*. Following [Hu *et al.*, 2021], we use the first four sessions for training, the last for testing, and randomly select 10% of the training set for validation.

MELD [Poria *et al.*, 2019] consists of video data from multi-party conversations in TV show "Friends", with utterances classified into seven emotion categories, i.e., *Neutral, Surprise, Fear, Sadness, Joy, Disgust, Anger*. We use the predefined splits for training and evaluation.

Dataset	Conversations			Utterances			Classes
	train	val	test	train	val	test	
IEMOCAP	120		31	5810		1623	6
MELD	1039	114	280	9989	1109	2610	7

Table 1: Statistics of the two benchmark datasets.

4.2 Baselines

We compare our proposed **MATCH** with ten baseline models including aggregation-based methods like **DiaRNN** [Majumder *et al.*, 2019] **CTNet** [Lian *et al.*, 2021] **CauAIN** [Zhao *et al.*, 2022] **CMERC** [Tu *et al.*, 2024b], graph-based methods like **DiaGCN** [Ghosal *et al.*, 2019] **MM-DFN** [Hu *et al.*, 2022] **AdaIGN** [Tu *et al.*, 2024a] **PCGNet** [Tu *et al.*, 2024c], and hypergraph-based methods like **M^{3}Net** [Chen *et al.*, 2023] **HAUCL** [Yi *et al.*, 2024].

4.3 Implementation Details

All experiments were conducted on an NVIDIA RTX A6000 GPU using the torch-geometric package. We used a batch size of 16 for both datasets. For IEMOCAP, the learning rate was set to 1e-4 with a dropout rate of 0.4, while for MELD, the learning rate was set to 5e-4 with a dropout rate of 0.3. Additional parameter settings are provided in Table 2.

Dataset	τ_1	η	τ_m	τ_c	τ_{emo}	η_2	L	K
IEMOCAP	0.5	1	0.4	0.5	0.5	3e-5	4	3
MELD	0.4	1	0.3	0.4	0.4	3e-5	4	4

Table 2: Hyperparameter settings on two datasets.

4.4 Results and Discussion

As shown in Table 3, **MATCH** outperforms existing approaches on both datasets. Specifically, ACC improves by 0.92% and W-F1 increases by 0.57% on IEMOCAP. Similarly, on MELD dataset, ACC improves by 0.31%, and W-F1 improves by 0.02%. Our method also achieves competitive F1 scores across most emotion categories, with notably optimal performance on "Neutral" and "Excited," surpassing state-of-the-art results.

An obvious drawback of aggregation-based approaches is the high coupling of contextual information across multiple time steps. While **CTNet** addresses this by introducing different GRU units and employing an attention mechanism to enhance discriminability, it still struggles to capture multiple complex conversational relationships. In contrast, **MATCH** improves upon these methods by calibrating entity information at both the contextual and modality levels, enhancing the quality of utterance representation and mitigating the interference of natural noise in the utterance, thus improving the subsequent fusion process.

Graph-based methods capture speaker and contextual interactions through edges, improving perception of complex relationships within conversations. Hypergraph-based models, particularly, offer superior emotional understanding due to the capacity to yield high-order relationships between multiple connected nodes. While **M^{3}Net** and **HAUCL** leverage this property of hypergraphs, **HAUCL** adaptively constructs hyperedges via a hypergraph variational autoencoder, reducing the misleading influence of redundant hyperedges on node information transfer. **M^{3}Net**, by contrast, enriches node-level information through multi-frequency decomposition. However, both methods primarily focus on node-level information, neglecting the role of hyperedges in contextual understanding. Our **MATCH** enhances hypergraph representations through line expansion, capturing both fundamental semantic expressions and deeper conversational patterns. The emotion-aligned learning further minimizes the information discrepancies between the two, enabling more effective emotion comprehension.

Interestingly, we find that **MATCH** outperforms **CauAIN** and **PCGNet**, which rely on external knowledge including personality traits and commonsense. This indicates that **MATCH** achieves comprehensive emotion understanding solely through multimodal data, without the need for external knowledge.

4.5 Ablation Study

We conduct ablation experiments on **MATCH** to validate each component's effectiveness.

Effect of Conversational Entity Calibration. As shown in Table 5, removing the entity calibration module significantly degrades **MATCH**'s performance, as it relies solely on initial features which contain coupled or redundant information, hindering reliable predictions. This unreliability is effectively alleviated by introducing entity calibration. While **M^{3}Net** attempts to mitigate coupling through multi-frequency decomposition, it does not fully distinguish the contributions of contextual and modality information, leading to suboptimal performance. We apply the proposed entity calibration strategy to several baselines, with the results in Table 4 showing improvements in all cases. This indicates that most baselines have limitations in utilizing multimodal features, and our entity calibration strategy provides a foundation for extracting deeper emotional cues.

Effect of Emotion-aligned Hypergraph Fusion. As **MATCH**'s core module, emotion-aligned hypergraph fusion is designed to capture high-order semantic cues and diverse dialogue relationships. Without it, **MATCH** struggles to perceive complex relationships, limiting its ability to facilitate effective cross-modal interactions essential for MERC. Furthermore, deeper conversational patterns are missed, preventing the accurate conveyance of similar emotions through contextual and modality relationships, and resulting in the loss of emotional pathways. This lack of deeper information severely impacts the model's ability to generate accurate emotion judgments.

Effect of Conversational Pattern Learning. We introduce line expansion to enhance the hypergraph's ability to represent high-order semantics, uncovering meaningful conversational patterns and emotional pathways, emphasizing the positive role of hyperedges in MERC. Without this, the node-focused propagation mechanism of hypergraph fails to effectively transmit hyperedge information, leading to redundant semantics being mistakenly propagated to connected nodes. Additionally, learning conversational patterns offers more contextualized emotional insights and effectively preserves implicit semantic transfer within conversations. Integrating conversational patterns with global semantic understanding through graph-level alignment enables a more comprehensive affective understanding.

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Acc	W-F1
★DiaRNN♯ [Majumder *et al.*, 2019]	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	76.97	47.69	-	20.41	50.92	-	45.52	60.31	57.66
★DiaGCN♯ [Ghosal *et al.*, 2019]	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	75.97	46.05	-	19.60	51.20	-	40.83	58.62	56.36
CTNet♮ [Lian *et al.*, 2021]	51.30	79.90	65.80	67.20	78.70	58.80	68.00	67.50	77.40	52.70	10.00	32.50	56.00	11.20	44.60	62.00	60.50
★MM-DFN♯ [Hu *et al.*, 2022]	42.22	78.98	66.42	**69.77**	75.56	66.33	68.21	68.18	77.76	50.69	-	22.93	54.78	-	47.82	62.49	59.46
♣CauAIN♮ [Zhao *et al.*, 2022]	-	-	-	-	-	-	-	67.61	-	-	-	-	-	-	-	-	65.46
CMERC♮ [Tu *et al.*, 2024b]	60.73	81.89	71.65	69.51	77.45	67.02	-	71.98	80.18	60.42	24.69	40.48	65.30	**32.31**	54.16	-	66.85
★M^{3}Net♭ [Chen *et al.*, 2023]	**61.27**	78.67	68.70	65.47	76.02	62.79	69.07	69.17	79.53	59.09	18.42	37.38	**65.32**	21.15	54.52	66.86	65.78
AdaIGN♮ [Tu *et al.*, 2024a]	53.04	81.47	71.26	65.87	76.34	67.79	-	70.74	79.75	60.53	-	43.70	64.54	-	56.15	-	66.79
★HAUCL♮ [Yi *et al.*, 2024]	53.57	82.04	68.61	66.44	75.60	68.23	70.30	70.27	80.01	59.85	21.95	36.72	63.79	29.31	55.54	67.62	66.23
♣★PCGNet♮ [Tu *et al.*, 2024c]	49.83	**82.70**	71.62	69.14	76.08	**70.98**	71.72	71.77	80.25	61.02	**25.88**	41.48	64.65	25.24	56.09	67.85	67.02
MATCH(Ours)	59.18	82.10	**74.44**	67.28	**79.67**	66.31	**72.64**	**72.55**	**80.38**	**61.34**	19.75	**41.82**	63.75	26.67	**56.98**	**68.16**	**67.04**

Table 3: Comparison of results against various MERC models. ★ denotes source code available. ♣ denotes the external knowledge is introduced in method. ♯, ♮ represents results from MM-DFN, and original papers, respectively. ♭ denotes results from out re-implementation.

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Acc	W-F1
DiaRNN♭	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	75.84	48.92	-	16.73	51.16	-	43.91	59.62	56.78
w/ Entity Calibration	33.98	83.90	56.50	60.39	73.09	60.48	64.02	63.04	76.87	46.10	-	14.88	53.87	-	44.09	60.27	57.26
DiaGCN♭	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	69.81	43.12	6.42	21.07	69.03	5.06	36.98	50.61	52.61
w/ Entity Calibration	34.21	70.48	60.89	63.47	72.83	60.48	64.26	63.55	71.10	48.10	5.13	30.07	51.16	2.44	44.14	54.14	55.67
MM-DFN♭	33.48	79.83	66.12	68.11	73.86	67.18	67.78	67.18	75.89	46.52	-	24.23	54.14	-	43.29	59.00	57.52
w/ Entity Calibration	37.19	78.50	66.50	69.72	77.46	67.84	68.76	68.38	76.88	47.59	-	30.90	52.09	-	41.75	60.54	58.13
M^{3}Net♭	61.27	78.67	68.70	65.47	76.02	62.79	69.07	69.17	79.53	59.09	18.42	37.38	65.32	21.15	54.52	66.86	65.78
w/ Entity Calibration	54.48	80.00	72.93	65.02	77.51	69.05	71.47	71.46	80.22	61.09	19.75	38.61	65.01	28.33	54.41	67.85	66.58
HAUCL♭	54.36	80.08	68.26	65.55	73.18	65.43	68.64	68.77	79.58	58.82	18.67	39.88	63.16	28.83	52.99	66.82	65.65
w/ Entity Calibration	50.77	78.19	70.07	64.22	71.48	68.21	68.88	68.79	79.37	60.20	25.00	40.00	64.44	25.00	53.95	66.93	66.05
PCGNet♭	46.03	83.33	69.38	65.24	77.24	64.35	69.44	69.25	76.23	58.10	26.67	43.87	64.50	26.85	54.22	63.49	64.75
w/ Entity Calibration	50.18	84.08	71.14	64.72	73.84	65.95	69.87	69.84	78.13	59.93	25.90	40.34	66.06	26.32	55.01	65.63	65.89

Table 4: Performance of various MERC methods with conversational entity calibration.

Methods	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
w/o Contextual Calibration	69.56	69.45	67.59	66.44
w/o Speaker Calibration	70.30	70.19	67.62	66.52
w/o Entity Calibration(full)	69.38	69.32	67.27	66.09
w/o Emotion-aligned Hypergraph Fusion	67.78	67.95	66.93	65.38
w/o Conversational Pattern Learning	70.43	70.64	67.13	66.14

Table 5: Ablation results for MATCH.

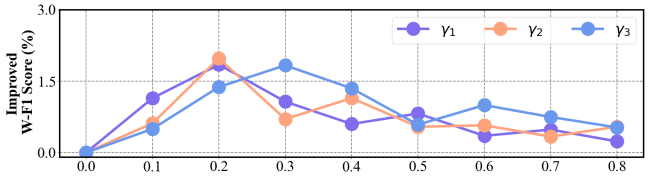

Figure 3: Improvement in W-F1 score of MATCH across different hyperparameters on IEMOCAP.

4.6 Hyperparameter Sensitivity Analysis

We analyze key hyperparameters γ_1, γ_2, γ_3, and the number of hypergraph (L) and line graph (K) layers on IEMOCAP validation set. Figure 3 demonstrates that the impact of these hyperparameters on the overall W-F1 results follows a trend of initial improvement, which then declines and stabilizes as their values increase. Excessively large parameter values reduce their effectiveness, indicating challenges in achieving an optimal balance between primary and auxiliary tasks. Figure 4 illustrates that performance improves initially, but declines as the number of hypergraph or line graph layers increases. An excessive number of layers fails to produce more meaningful emotion representations.

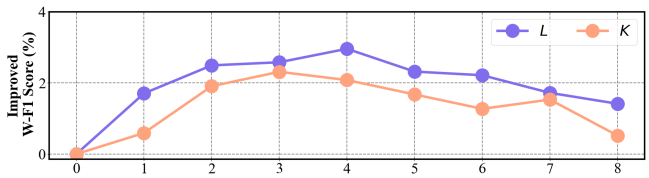

Figure 4: Improvement in W-F1 score of MATCH with varying hypergraph and line graph layers on IEMOCAP.

5 Conclusion

In this paper, we introduce **MATCH**, a hypergraph-based framework for MERC. **MATCH** enhances utterance representations by calibrating dialogue entities at context and modality perspectives. It captures high-order substructural interactions in the hypergraph through line expansion, extracting deep emotional cues while preserving fundamental semantics, enabling effective emotional pathway learning. Besides, emotion-aligned learning improves knowledge transfer at hyperedge and graph levels. Experiments demonstrate that **MATCH** outperforms state-of-the-art methods on two datasets without relying on external knowledge, fully leveraging the potential of hypergraphs in MERC.

Acknowledgments

This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C03262), and by the National Natural Science Foundation of China (No. U21A20473, No. 62172370, No. 62176244, No. 62032022, No. U23A20388, No. 62320106007).

Contribution Statement

Ming Li (email: mingli@zjnu.edu.cn) is the corresponding author of this work.

References

[Bertero *et al.*, 2016] Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1042–1047, 2016.

[Busso *et al.*, 2008] 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.

[Chen *et al.*, 2023] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10761–10770, 2023.

[Chen *et al.*, 2024] Yin Chen, Xiaoyang Wang, and Chen Chen. Hyperedge importance estimation via identity-aware hypergraph attention network. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 334–343, 2024.

[Duan *et al.*, 2024] Junwei Duan, Jiaqi Xiong, Yinghui Li, and Weiping Ding. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 112:102536, 2024.

[Feng *et al.*, 2019] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3558–3565, 2019.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 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*, pages 154–164, 2019.

[Hu *et al.*, 2021] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 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*, pages 5666–5675, 2021.

[Hu *et al.*, 2022] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7037–7041, 2022.

[Joshi *et al.*, 2022] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram 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.

[Ju *et al.*, 2024] Wei Ju, Zhengyang Mao, Siyu Yi, Yifang Qin, Yiyang Gu, Zhiping Xiao, Yifan Wang, Xiao Luo, and Ming Zhang. Hypergraph-enhanced dual semi-supervised graph classification. In *Forty-first International Conference on Machine Learning*, 2024.

[Li *et al.*, 2023a] 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.

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

[Li *et al.*, 2024] Ming Li, Siwei Zhou, Yuting Chen, Changqin Huang, and Yunliang Jiang. Educross: Dual adversarial bipartite hypergraph learning for cross-modal retrieval in multimodal educational slides. *Information Fusion*, 109:102428, 2024.

[Li *et al.*, 2025a] Ming Li, Yukang Cheng, Lu Bai, Feilong Cao, Ke Lu, and Jiye Liang. EduLLM: Leveraging large language models and framelet-based signed hypergraph neural networks for student performance prediction. 2025.

[Li *et al.*, 2025b] Ming Li, Yujie Fang, Yi Wang, Han Feng, Yongchun Gu, Lu Bai, and Pietro Liò. Deep hypergraph neural networks with tight framelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18385–18392, 2025.

[Li *et al.*, 2025c] Ming Li, Yongchun Gu, Yi Wang, Yujie Fang, Lu Bai, Xiaosheng Zhuang, and Pietro Liò. When hypergraph meets heterophily: New benchmark datasets and baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18377–18384, 2025.

[Li *et al.*, 2025d] Ming Li, Jiandong Shi, Lu Bai, Changqin Huang, Yunliang Jiang, Ke Lu, Shijin Wang, and Edwin R Hancock. Frameerc: Framelet transform based multimodal graph neural networks for emotion recognition in conversation. *Pattern Recognition*, 161:111340, 2025.

[Lian *et al.*, 2021] 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.

[Lu *et al.*, 2024] Nannan Lu, Zhiyuan Han, and Zhen Tan. A hypergraph based contextual relationship modeling method for multimodal emotion recognition in conversation. *IEEE Transactions on Multimedia*, pages 1–13, 2024.

[Mai *et al.*, 2020] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on artificial intelligence*, pages 164–172, 2020.

[Majumder *et al.*, 2019] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, pages 6818–6825, 2019.

[Millán *et al.*, 2025] Ana P Millán, Hanlin Sun, Lorenzo Giambagli, Riccardo Muolo, Timoteo Carletti, Joaquín J Torres, Filippo Radicchi, Jürgen Kurths, and Ginestra Bianconi. Topology shapes dynamics of higher-order networks. *Nature Physics*, 21:353–361, 2025.

[Mittal *et al.*, 2020] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020.

[Poria *et al.*, 2019] 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.

[Shi *et al.*, 2025] Jiandong Shi, Ming Li, Yuting Chen, Lixin Cui, and Lu Bai. Multimodal graph learning with framelet-based stochastic configuration networks for emotion recognition in conversation. *Information Sciences*, 686:121393, 2025.

[Tellamekala *et al.*, 2023] 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.

[Tu *et al.*, 2024a] 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*, pages 19089–19097, 2024.

[Tu *et al.*, 2024b] Geng Tu, Feng Xiong, Bin Liang, Hui Wang, Xi Zeng, and Ruifeng Xu. Multimodal emotion recognition calibration in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9621–9630, 2024.

[Tu *et al.*, 2024c] 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.

[Wang *et al.*, 2023] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. In *Proceedings of the Neural Information Processing Systems.*, 2023.

[Wang *et al.*, 2024] Keke Wang, Yu Zhu, Xiaoying Wang, Jianqiang Huang, and Tengfei Cao. Heterogeneous hypernetwork representation learning with hyperedge fusion. *IEEE Transactions on Computational Social Systems*, 11(6):7646–7657, 2024.

[Yang *et al.*, 2022] Chaoqi Yang, Ruijie Wang, Shuochao Yao, and Tarek F. Abdelzaher. Semi-supervised hypergraph node classification on hypergraph line expansion. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2352–2361, 2022.

[Yi *et al.*, 2024] Zijian Yi, Ziming Zhao, Zhishu Shen, and Tiehua Zhang. Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4341–4348, 2024.

[Zadeh *et al.*, 2017] 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.

[Zadeh *et al.*, 2018] 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*, pages 5634–5641, 2018.

[Zhang *et al.*, 2024] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pages 1807–1817, 2024.

[Zhao *et al.*, 2022] Weixiang Zhao, Yanyan Zhao, and Xin Lu. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4524–4530, 2022.

[Zhao *et al.*, 2024] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36, 2024.