Scene-Speaker Emotion Aware Network
Dual Network Strategy for Conversational Emotion Recognition

Li, Bingni ; Gu, Yu ; Li, Chenyu ; Zhang, He ; Liu, Linsong ; Lin, H.X.; Wang, Shuang

DOI
10.3390/electronics14132660

Publication date
2025

Document Version
Final published version

Published in
Electronics (Switzerland)

electronics

MDPI

Article

Scene-Speaker Emotion Aware Network: Dual Network Strategy for Conversational Emotion Recognition

Bingni Li [1], Yu Gu [1,*], Chenyu Li [1], He Zhang [2], Linsong Liu [1], Haixiang Lin [3] and Shuang Wang [1]

[1] School of Artificial Intelligence, Xidian University, Xi'an 710126, China; 20009200090@stu.xidian.edu.cn (B.L.); 22171214755@stu.xidian.edu.cn (C.L.); 23171214603@stu.xidian.edu.cn (L.L.); shwang@mail.xidian.edu.cn (S.W.)
[2] School of Journalism and Communication, Northwest University, Xi'an 710127, China; zhanghe@nwu.edu.cn
[3] Delft Institute of Applied Mathematics, Technische Universiteit Delft, CN 2628 Delft, The Netherlands; h.x.lin@tudelft.nl
* Correspondence: guyu@xidian.edu.cn

Abstract

Incorporating external knowledge has been shown to improve emotion understanding in dialogues by enriching contextual information, such as character motivations, psychological states, and causal relations between events. Filtering and categorizing this information can significantly enhance model performance. In this paper, we present an innovative Emotion Recognition in Conversation (ERC) framework, called the Scene-Speaker Emotion Awareness Network (SSEAN), which employs a dual-strategy modeling approach. SSEAN uniquely incorporates external commonsense knowledge describing speaker states into multimodal inputs. Using parallel recurrent networks to separately capture scene-level and speaker-level emotions, the model effectively reduces the accumulation of redundant information within the speaker's emotional space. Additionally, we introduce an attention-based dynamic screening module to enhance the quality of integrated external commonsense knowledge through three levels: (1) speaker-listener-aware input structuring, (2) role-based segmentation, and (3) context-guided attention refinement. Experiments show that SSEAN outperforms existing state-of-the-art models on two well-adopted benchmark datasets in both single-text modality and multimodal settings.

Keywords: attention mechanism; commonsense knowledge; emotion recognition in conversation; multimodal fusion

Academic Editor: M-Tahar Kechadi

Received: 23 April 2025
Revised: 16 June 2025
Accepted: 25 June 2025
Published: 30 June 2025

Citation: Li, B.; Gu, Y.; Li, C.; Zhang, H.; Liu, L.; Lin, H.; Wang, S. Scene-Speaker Emotion Aware Network: Dual Network Strategy for Conversational Emotion Recognition. *Electronics* **2025**, *14*, 2660. https://doi.org/10.3390/electronics14132660

1. Introduction

Emotion recognition has been a prominent research area in natural language processing over time [1]. With the rapid proliferation of social media and online platforms, an unprecedented volume of conversational data has become available for computational analysis [2]. This has fueled increasing interest in Emotion Recognition in Conversation (ERC), which is essential for various downstream applications, such as emotion-driven chatbots [3], automated customer service [4,5], and sentiment analysis on social media platforms [6,7].

Recent ERC studies have explored various strategies for modeling speaker-specific emotional states. Some employ recurrence-based methods for speaker modeling [8–11], utilizing separate recurrent networks to distinguish between speakers and listeners. Others introduce Graph Convolutional Networks (GCNs) [12] to represent utterances and speaker relationships [13–17] by effectively capturing internal emotional inertia and speaker interactions. However, to capture causal information in context, these methods require extracting

utterance-level features early on, which include the identity information of different speakers. This shifts the model's focus away from the emotional dynamics of the specific speaker and, thus, persisting in using these features may lead to confusion.

In addition, some other ERC methods enhance emotional modeling by integrating external knowledge to better address complex speaker-related factors, such as psychological states and speaking motives [18,19]. Notably, classic methods such as COSMIC [20] and TodKat [21] utilize structured commonsense knowledge bases like ConceptNet [22] and ATOMIC [23], or automatically constructed knowledge graphs based on them (e.g., COMET [24]), to enhance emotional reasoning. Results demonstrate that these knowledge-integrated approaches significantly improve the model's ability to capture nuanced emotional dynamics and uncover implicit speaker intentions, thereby enhancing overall emotion understanding in dialogue.

While a few recent methods attempt to extract additional commonsense knowledge from the inherent knowledge of large language models (LLMs) [25] or reasoning networks [26], the majority still rely on structured commonsense knowledge bases as their primary external source. These knowledge graphs typically offer nine types of relational information. To determine the most effective combination, existing methods often evaluate model performance under various configurations. However, the selected relation types vary considerably across studies. For instance, some methods, such as MKIN-MCL [27], focus on six specific relation types, while CDEA [28] and LECM [29] select six relations based on subject–object roles in event semantics. In contrast, other methods prioritize fewer but more relevant relations—COFFEE [30] selects two based on their correlation with emotion labels, and TG-ERC [31] utilizes three associated with psychological states. However, this trial-and-error process is inefficient and overlooks the potential unreliability of generated external knowledge, often noisy or factually incorrect, which can introduce harmful bias into the model. While some methods aim to suppress unnecessary commonsense knowledge, for instance, the CKCL method [32] leverages contrastive learning to reduce reliance on external knowledge when it aligns with contextual predictions, they fail to address cases where unreliable knowledge misleads otherwise accurate predictions. No existing approach effectively filters and aligns external commonsense knowledge at the utterance level, which undermines its ability to serve as a positive and contextually appropriate supplement.

To improve the efficiency of leveraging multimodal information and external commonsense knowledge in emotion recognition in conversation (ERC), two key challenges must be addressed: (i) Modeling a speaker's emotional state based on global utterance-level features may introduce noise from mixed speaker identity information, especially in multi-speaker dialogues; (ii) External commonsense knowledge can be unreliable and noisy. There is a lack of effective and universally applicable filtering mechanisms to ensure its relevance and accuracy.

We propose the Scene-Speaker Emotion Aware Network (SSEAN), which employs dual parallel recurrent networks to model both global context and individual speaker emotions. This approach categorizes input information to capture global-level and speaker-specific emotions independently. In multi-turn dialogues, this helps the model capture the continuity and correlation of each speaker's emotional state across turn transitions. It also enhances the model's focus on the utterance features that are truly relevant to emotion by alleviating the noise introduced by abrupt speaker identity changes. Furthermore, we introduce a novel dynamic screening module to enhance commonsense knowledge across three levels, including: (1) Structuring compound single- and dual-sentence inputs based on speaker continuity to model speaker-listener dynamics, (2) segmenting and organizing generated commonsense knowledge into speaker and listener paragraphs at

the output level, and (3) dynamically filtering role-specific commonsense features using a context-guided attention mechanism. Experiments on the IEMOCAP and MELD datasets demonstrate that our model achieves, and in some cases surpasses, SOTA-level performance in both single-text modality and multimodal settings, underscoring the effectiveness of our approach.

2. Related Work

ERC Methods Focused on Speaker Modeling: The interactive conversational memory network (ICON) [8] pioneered the use of distinct memory networks to handle the interactions between speakers in dyad ic dialogues. The model first utilizes distinct gated recurrent unit (GRU) modules to capture speaker-specific contextual representations for each utterance, which are then integrated through global context modeling. Dialoguernn [9] also added two new GRUs to differentiate the impact of new utterances on speakers and listeners, enabling the model to extend to multi-party dialogues. Inspired by this, our model also employs GRUs to update the emotional states of different speakers. In DialogueGCN [14], directed graph network structures were introduced into dialogue emotion recognition to better model the interactions between speakers and the emotional inertia within individual speakers. They further modeled the speaker-level context by establishing a graph structure of adjacent utterances to the target utterance. I-GCN [13] designed two GCNs to process semantic information at the utterance level and relationship information at the speaker level, respectively, and used an incremental graph structure to capture temporal change information. Concurrently, the Directed Acyclic Graph Network for Conversational Emotion Recognition (DAG-ERC) [17] utilized speaker information and utterance position information to construct a directed acyclic graph neural network to model the dialogue context, enhancing the model's ability to capture long and short-term sequential information. To address the limitation of recurrent networks in simultaneously modeling dialogue structure and speaker information due to their sequential nature, we employ parallel recurrent networks to capture these two types of information separately.

ERC Methods Focused on Multimodality: Many other approaches focus on multimodal fusion, making full use of effective information across different modalities through the comprehensive application of cross-attention mechanisms and feature decoupling. The multimodal Dynamic Fusion Network (MM-DFN) [16] employs a novel graph-based dynamic fusion module to capture the dynamics of contextual information across different semantic spaces, significantly advancing the state of multimodal emotion recognition in conversations. CFN-ESA [11] incorporates a cross-modal fusion network with emotion-shift awareness, utilizing the textual modality as the primary source. It employs a novel cross-modal encoder module to fully extract complementary and associative information from multimodal data. Li et al. (2022) [10] made improvements in the feature extraction approach. To ensure that features extracted from each modality are more focused on emotional information, they proposed the Emoformer module for extracting emotion vectors to capture the subtle changes in emotions across different modalities, achieving significant performance improvements on two benchmark datasets. Our model draws on the method of extracting the emotional tendencies of each modality using variants of the transformer encoder.

ERC Methods Focused on Commonsense Knowledge: Emotion recognition enhanced by external knowledge mainly relies on two well-established commonsense knowledge bases. The first is ConceptNet [22], which captures commonsense concepts and relationships as a semantic network, covering various aspects of everyday life. The second is ATOMIC [23], centered on events rather than entities, achieving human-competitive results in If-Then reasoning tasks. Building on ATOMIC and ConceptNet, COMET [24], which can

automatically construct knowledge graphs, demonstrates the potential to understand and predict emotions, laying a foundation for incorporating commonsense knowledge into emotion recognition. COSMIC [20] and TokDat [21] leverage the COMET model to incorporate commonsense knowledge, thereby enhancing performance on dialogue emotion recognition. COSMIC [20] is a model that uses commonsense knowledge to model various hidden emotional influence factors in conversations, significantly improving the identification of complex emotions. Inspired by COSMIC, we also incorporate commonsense knowledge to model the emotional states of speakers and the complex influences among interlocutors. TokDat [21] combined a topic-augmented language model with commonsense statements, introducing them into a Transformer-based emotion detection model, achieving excellent accuracy. To acquire interpretable and relatively high-quality commonsense knowledge tailored for ERC tasks, we likewise adopt COMET as our external knowledge source. CKCL [32] is a contrastive learning framework designed to determine whether external knowledge is necessary for understanding utterance emotions, thereby avoiding the blind incorporation of knowledge that could hinder model training. It generates pseudo-labels based on the consistency between the original model prediction and the predictions obtained by masking either the context or the knowledge. However, this approach overlooks the fact that, even when the predictions are inconsistent, the incorporated commonsense knowledge may still act as noise and negatively affect the model.

Existing commonsense knowledge in ERC is typically centered around the speaker as the subject, inherently relying on speaker-specific information and requiring intra-speaker sequential modeling. Meanwhile, the generated commonsense knowledge may contradict the fact, which may introduce noise and hinder understanding of speaker's actual state. However, existing ERC methods rarely filter such knowledge effectively. Our proposed model employs parallel recurrent networks to separately capture scene-level and speaker-level emotional information, enabling better use of commonsense knowledge for speaker modeling and emotional interaction understanding, while preserving global context comprehension. Additionally, a multi-stage attention-based filtering module is introduced to improve the quality of generated commonsense knowledge and identify potential noise.

3. The Proposed Method

We propose the Scene-Speaker Emotion Aware Network (SSEAN), a unified framework for emotion recognition in conversation that jointly models global context and speaker-specific emotional dynamics. The SSEAN employs dual parallel recurrent networks to separately capture global scene-level context and individual speaker emotions, effectively preserving emotional continuity across multi-turn dialogues while reducing noise from abrupt speaker identity shifts. To further enhance reasoning with external knowledge, we design a three-level dynamic commonsense screening module that improves knowledge quality and relevance. The overall architecture of our model is illustrated in Figure 1.

3.1. Problem Definition

Formally, given a conversation C consists of a series of utterances $C = [u_1, u_2..., u_N]$, $u_i = \{u_i^t, u_i^a, u_i^v\}$ where N is the number of utterances in the conversation, and u_i denotes the i^{th} utterance in the conversation, which contains the representations of two modalities u_i^t (text) , u_i^a (audio) and u_i^v (visual). For the conversation C, speakers $P = \{p_1, p_2, ..., p_M\}$ participate in the conversation, where M is the number of participants, and a function $p_j = S(u_i), i \in \{1, 2, ..., N\}, j \in \{1, 2, ..., M\}$ is defined for obtaining the speaker is defined for obtaining the speaker p_j of the utterance u_i. The objective of emotion recognition in conversation is to accurately predict the emotion label y_i for each utterance u_i in the given

conversation C from a predefined set of emotion labels $Y = [y_1, y_2..., y_k]$, where k is the number of labels.

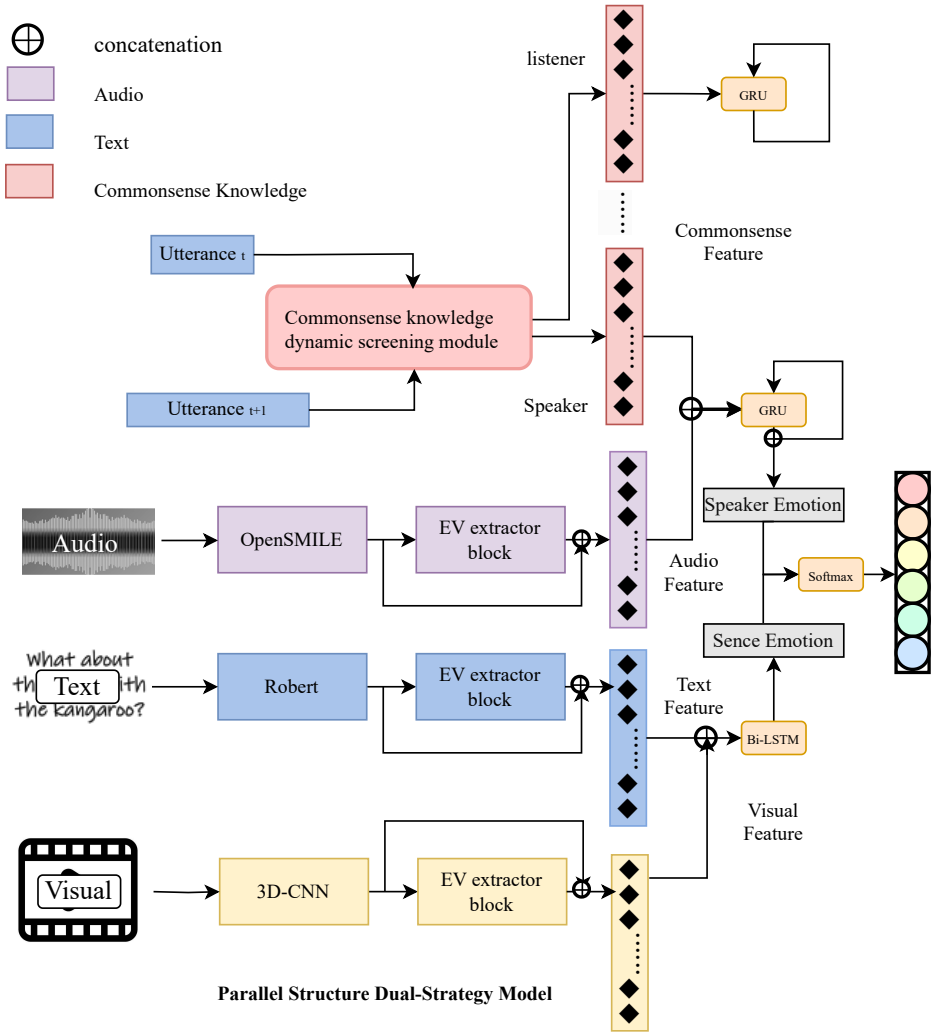

Figure 1. Framework illustration of Muti-model extended SSEAN.

3.2. Single Modality Feature Extraction

3.2.1. Raw Feature Extraction

First, this study utilizes pre-trained models and tools to extract raw features from individual modalities.

Textual Features

The RoBERTa Large model [33] is utilized for extracting textual representations at the utterance level. BPE tokenized utterances are fed into the model, where the encoder module of RoBERTa is employed for feature extraction, and the decoder module is omitted. In alignment with COSMIC [20], the outputs of the last four hidden layers are averaged to enrich the features with maximal information, resulting in raw text features for enhanced context modeling.

Audio Feature

Following several previous studies [9,34], this paper uses the standard sets ComParE 2016 from the OpenSMILE [35] as a profile for the initial processing of the audio data. ComParE 2016 is the feature set required by The INTERSPEECH 2016 ComParE Challenge, which contains 6373 static features obtained by computing various functions on LLD and is

suitable for a variety of downstream tasks including emotion recognition. In this paper, given that the feature dimensions directly output by OpenSMILE are relatively large, a fully connected layer is utilized to reduce the dimensionality of the features, yielding condensed audio raw features.

Visual Feature

Similar to the audio modality, to ensure a fair comparison, we also adopt the approach used in previous studies and employ a 3D-CNN for visual feature extraction. 3D-CNN extracts facial features by leveraging 3D convolutional layers and 3D pooling layers, capturing information across both spatial and temporal dimensions, and is commonly used for facial expression recognition [36]. This is highly relevant to conversational emotion recognition.

After this step, we represent the features of the utterance u_i as $U_i = [U_i^t, U_i^a, U_i^v]$.

3.2.2. Emotion Vector Extraction

To enhance the Scene-Speaker Emotion Aware Network's emphasis on the continuity and variation of emotional information while minimizing the impact of emotion-irrelevant noise, this study draws inspiration from EmoCaps [10]. It incorporates a structure similar to the Transformer encoder for processing each modality, facilitating the nuanced extraction of emotional features, which has been demonstrated to be feasible in EmoCaps. In the Emotion Vector extraction block, as illustrated in Figure 2, two encoder modules were modified and merged, with the subsequent feed-forward network being replaced by a multi-layer perceptron. Since the self-attention mechanism has a good ability to capture global information from a long sequence, this paper uses it to obtain utterance-level emotion information further. At last, the block aggregates the emotion information through the multilayer perceptron to reduce the feature dimensions and obtain more representative unimodal utterance-level emotion vectors. Since the input single-utterance raw features are unrelated to the dialogue-level context, the emotional vectors extracted at this stage are also independent of the dialogue-level context.

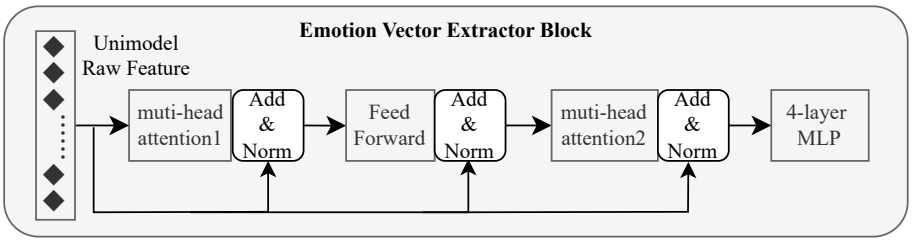

Figure 2. Emotion Vector block.

For a given utterance u_i, the process of computing the utterance-level emotion vector E_i through the Emotion Vector extraction block can be expressed as follows:

$$\begin{aligned} A1_i^x &= LayerNorm(U_i^x + soft\max(\frac{U_i^x W_q U_i^{x^T} W_k^T}{\sqrt{d}})U_i^x W_v) \\ F_i^x &= LayerNorm(U_i^x + RELU(A_i^x W + b)) \end{aligned} \tag{1}$$

where $Ui^x, x \in \{t, a\}$is the text or audio component of U_i, W_q, W_k, W_v, W is the learnable parameter matrix, and b is the bias parameter. For each modality component, we stack two identical self-attention–residual–LayerNorm blocks. and then fed into a multilayer perceptron:

We simplify the feedforward network, which usually consists of two fully connected layers with ReLU activation functions, into one layer, retaining the nonlinear transformation

capability while using the multilayer perceptron as a substitution to further improve the nonlinear fitting capability and reducing the feature dimension.

Finally, we concatenation the utterance-level emotion vector $E_i = [E_i^t, E_i^a, E_i^v]$ with the raw feature $U_i = [U_i^t, U_i^a, U_i^v]$ based on modality to obtain the final representation of the new discourse u_i unimodal feature:

$$\begin{aligned} TF_i &= E_i^t \oplus U_i^t \\ AF_i &= E_i^a \oplus U_i^a \\ VF_i &= E_i^v \oplus U_i^v \end{aligned} \tag{2}$$

3.3. Dynamic Screening of Commonsense Knowledge

For a given utterance, we take it as input and use COMET [24] trained on ATOMIC, a knowledge generation model, as the only source to acquire the corresponding commonsense knowledge related to the speaker's emotion state. ATOMIC is an event-centered knowledge graph that allows for the execution of the corresponding inference task based on the 9 if-then relation types identified as (i) xIntent, (ii) xNeed, (iii) xAttr, (iv) xEffect, (v) xWanted, (vi) xReact, (vii) oEffect, (viii) oWant, and (ix) oReact [23].

Referring to existing work that enhances ERC with commonsense knowledge [20,21], this paper excludes the relation types xNeed, xWant and oWant because they are predictions of character actions before and after the event. Whereas in the dialogue dataset, considering that each dialogue lasts for a shorter period, we do not assume that more actions take place during the conversation. Yet, there is still a controversy about the role of the remaining part of the relationship types for sentiment recognition, existing work [20,21] have experimentally sifted the relation types used in the model species. It has been observed that incorporating additional relation types into the model results in a decline in model performance. However, even within the same dialogue, the applicability of relation types to utterances can vary. For some utterances, all relation categories can provide valid commonsense knowledge, while for other utterances, only some of the relation categories may be able to provide valid commonsense knowledge. Thus, the manual selection method can not make the best use of commonsense knowledge. In this paper, we use the remaining six relational categories for our experiments. The usage of relationship types in related work is shown in Table 1.

Table 1. Relation types used in related work.

Related Work	xInt	xRea	xAtt	xEff	oRea	oEff
COSMIC	✓	✓		✓	✓	✓
TODKAT	✓	✓			✓	
SSEAN	✓	✓	✓	✓	✓	✓

In addition, as COMET [37] is built upon a pre-trained GPT-2 model, the knowledge it generates tends to be diverse and includes multiple plausible alternatives. Additionally, the generated commonsense knowledge is inevitably speculative, based on current circumstances, and its reliability requires validation by subsequent factual developments. Directly using raw dialogue utterances as input for COMET further exacerbates these challenges, introducing additional issues that compromise the quality and relevance of the generated knowledge. These speculative, low-quality pieces of commonsense knowledge should not be directly adopted by the model, as they could potentially mislead it.

Therefore, to ensure that the model acquires sufficient and reliable commonsense knowledge, we propose a three-tier dynamic filtering module to refine and enhance the quality of commonsense knowledge. The main structure is shown in Figure 3.

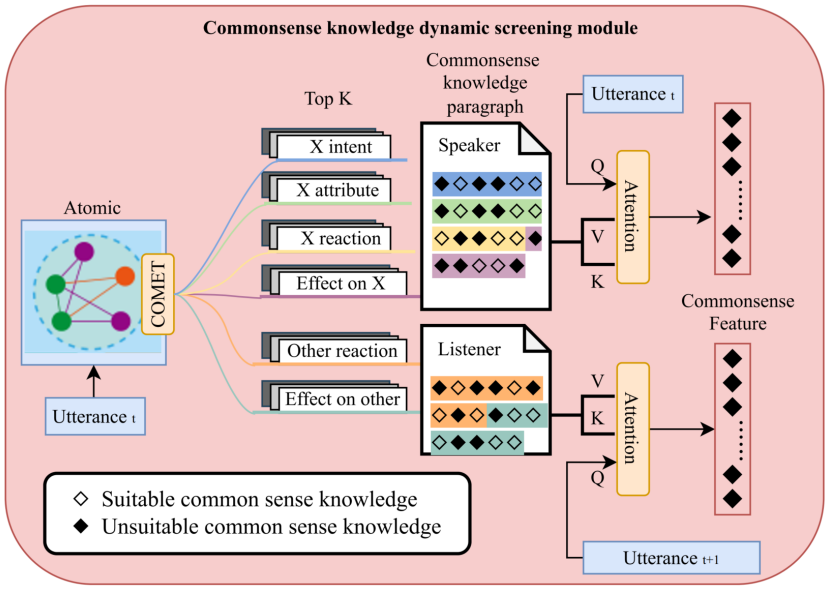

Figure 3. Commonsense knowledge dynamic screening moudle.

First, most previous studies directly use dialogue utterances as input, which presents two key problems. The first issue is that COMET's training data consists of descriptive statements with subject–verb–object structures, whereas in dialogues, the speaker and listener are often omitted from the utterances. The second issue arises from the segmentation criterion of dialogue datasets, which is typically based on punctuation marks such as periods. As a result, some utterances contain limited information, and directly inputting them into COMET may generate meaningless or even ambiguous knowledge.

To address these issues at the input level, we adopt a strategy that combines sentence completion with both single- and dual-sentence inputs. Specifically, for each utterance, we complete the missing subjects and verbs following the format of COMET's training data. Additionally, when a change in speaker occurs, we use the listener of the next utterance as the object and incorporate the subsequent utterance as a response to further enhance the input, thereby forming a dual-sentence input structure. As shown in Algorithm 1.

For each of the 6 relation types containing potentially valid information, the top k most plausible pieces of knowledge are generated in text form as candidates.

At the output level, we implement an initial filtering step. Our experiments reveal that the generated outputs often contain meaningless words or symbols, such as "none", ".", " ", "y", "x", and "n/a". We first eliminate such outputs and, for utterances that fail to produce meaningful commonsense knowledge, we apply padding using [pad] as a fallback mechanism. After that, by adding descriptive sentence components or subjects, we integrate all candidate knowledge into two general knowledge paragraphs according to subject differences. The detailed methodology is provided in Algorithm 2. The final filtered knowledge output from these steps is treated as input features for our end-to-end trainable model. Therefore, the inclusion of these algorithms does not break the end-to-end trainability of the SSEAN architecture.

To further ensure the reliability of commonsense knowledge across different subjects, we apply an additional filtering and guidance process to the generated commonsense knowledge paragraphs. Specifically, for the commonsense knowledge extracted from the current discourse, we identify the current discourse (for the speaker's knowledge paragraph) and the subsequent discourse (for the listener's knowledge paragraph) as valid facts. Commonsense knowledge that exhibits greater similarity to these valid facts is considered more reliable.

Algorithm 1: Speaker-listener-aware input structuring.

Dialogue $D = [u_1, u_2, \ldots, u_n]$ with speaker annotations Transformed input sequences $\hat{D} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n]$
Initialize $\hat{D} \leftarrow []$
for $i = 1$ *to* n **do**
  $u \leftarrow u_i$
  $s \leftarrow \text{speaker}(u)$
  $u \leftarrow \text{CompleteSubjectVerb}(u)$ ▷ Complete missing subject/verb using COMET format
  if $i < n$ ***and*** $speaker(u_{i+1}) \neq s$ **then**
    $v \leftarrow u_{i+1}$
    $l \leftarrow \text{speaker}(v)$
    $v \leftarrow \text{CompleteSubjectVerb}(v)$
    $v \leftarrow \text{InsertListenerAsObject}(v, l)$
    $\hat{u} \leftarrow u$ + "after that" + v
  else
    $\hat{u} \leftarrow u$
  end
  Append $\hat{u}$ to $\hat{D}$
end
return $\hat{D}$

Algorithm 2: Generate speaker and listener paragraphs.

Input: topk, model COMET, speaker identity S(·),
Descriptive_Components =
{xInt: 'wanted to', xRea: 'will feel', xAtt: 'is seen as',
xEff: 'will', oRea: 'will feel', oEff: 'will'}
Output: SpeakerParagraph, ListenerParagraph

for *all* $c \in \{xInt, xRea, xAtt, xEff, oRea, oEff\}$ **do**
  filtered = [];
  result = COMET.getCKsequence(, c, topk);
  for *all event in result* **do**
    if *event is not 'none', not '.', and not an empty string* **then**
      Add the event to filtered;
    end
  end
  if $c \in \{xInt, xRea, xAtt, xEff\}$ **then**
    if *length of filtered is 1* **then**
      sentence = descriptive_components[c] + filtered[0] + ".";
    else
      sentence = descriptive_components[c] + concatenate all event in filtered with ' and ' + ".";
    end
    SpeakerParagraph = SpeakerParagraph + " " + sentence;
  else
    if *length of filtered is 1* **then**
      sentence = descriptive_components[c] + filtered[0] + ".";
    else
      sentence = descriptive_components[c] + concatenate all event in filtered with ' and ' + ".";
    end
    ListenerParagraph = ListenerParagraph + " " + sentence;
  end
end
return *SpeakerParagraph, ListenerParagraph*

Using the same encoder as for the raw textual modality, we obtain utterance features f_s, f_l (current speaker and next-utterance listener) and the corresponding candidate knowl-

edge features C_s, C_l from COMET. For each role $n \in \{s, l\}$, queries, keys and values are computed as

$$Q_n = f_n W_q^n, \qquad K_n = C_n W_k^n, \qquad V_n = C_n W_v^n,$$

where W_q^n, W_k^n, W_v^n are learnable parameter matrices. The most informative and reliable knowledge representation is obtained through a single cross-attention layer followed by a position-wise feed-forward network:

$$CF_i^n = \mathrm{FFN}\Big(\mathrm{softmax}\big(\tfrac{Q_n K_n^\top}{\sqrt{d}}\big)\, V_n\Big), \qquad n \in \{s, l\}.$$

where the FNN is a two-layer feedforward network containing a ReLU activation layer. Since the attention mechanism is used as a filter here, no residual structure is added. The obtained output CF_i^s, CF_i^l is the speaker commonsense feature vector and listener commonsense feature vector for the given discourse u_i.

3.4. Dual-Strategy Framework

In this paper, we propose the Scene-Speaker Emotion Aware Network (SSEAN), a dialogue emotion recognition framework that employs dual-strategy parallel modeling to distinguish between the global conversational context and speaker-specific context, enabling the simultaneous utilization of multimodal and multi-source information.

Conversational Emotion recognition differs from general emotion recognition tasks in that it is difficult to make correct judgments about emotion by focusing only on utterance-level features. A significant amount of information is contained within the dialogue-level context and the multi-turn interactions among speakers. The components that can reflect the emotions of the utterances require selection and extraction through effective modeling methods.

Fundamentally, based on contextual relevance, we categorize emotional information within dialogues into two mutually exclusive types. The first type exhibits contextual relevance at the global dialogue level but loses this relevance within the same speaker, which involves the textual modality of utterances. Conversely, the second type includes commonsense knowledge about speaker states and the audio modality of utterances, showing contextual relevance within the same speaker while containing a lot of redundant speaker information at the global dialogue level. In this paper, this type involves textual modality of utterances and commonsense knowledge related to the speaker's state. To minimize the introduction of redundant information for these two types, we adopt different modeling strategies. We propose a parallel structure designed to capture the dialogue-level context, the emotional states of speakers, and multi-turn interactions between speakers independently. This architecture ensures that information about each dialogue participant remains distinct, minimizing redundant speaker-related data. The specific structure is shown in Figure 1.

A significant portion of information in dialogue is often embedded within long-term dependencies. Therefore, global-level contextual relationships can help the model better comprehend the overall progression and state of events throughout the conversation. By maintaining and updating dialogue history across multiple turns, the model gains a deeper understanding of the emotional tone underlying the conversation. Consequently, we refer to the features extracted based on global contextual information as global vectors. The Scene Emotion Vector is utilized to aid the model in understanding the continuity of emotions between adjacent utterances, such as being consistently neutral or negative throughout a particular paragraph. In this paper, we use a Bi-directional Long Short-Term Memory (Bi-LSTM) network [38] to model the global conversational context and extract the Scene Emotion Vector for each utterance from both video and textual modality features.

$$E_i^{scene} = LSTM(TF_i \oplus VF_i^s) \tag{3}$$

Throughout the process, emotional information can naturally be categorized according to the participants of the conversation. For the same event, the identity of the conversation participants might have a significant impact on the emotion of the utterances, which is particularly evident in multi-participant dialogues. Thus, we refer to the features extracted based on the state of conversation participants as Speaker Emotion Vectors. The update mechanism of the Speaker Scene Vector is utilized to help the model understand the inertia of emotions within the same speaker and the emotional interactions triggered between different speakers by the utterances. Due to the model's structure, information centered on different conversation participants remains independent, preventing cross-interaction. This design minimizes the introduction of speaker-related redundant information and mitigates its negative impact on model performance. To ensure this, we automatically assign independent GRU networks [39] to each speaker based on speaker labels, allowing the model to update the emotional states of individual speakers and obtain a Speaker Emotion Vector for each utterance.

$$E_i^{\text{speaker}} = e_i^{S(u_i)} = GRU_s(e_{i-1}^{S(u_i)}, (AF_i \oplus CF_i^s)) \tag{4}$$

In addition with the help of commonsense knowledge, we have also modeled the interaction between different speakers. Carrying on from the work in the previous section, we extracted the commonsense knowledge features by obtaining two commonsense knowledge feature vectors CF_i^s, CF_i^l whose subjects are the speaker and the listener of utterance u_i, respectively, where CF_i^l contains the current speaker's influence on the listener's emotion state, which is used to update the listener's emotion state.

This maintenance of the listener's state effectively captures the interaction dynamics between speakers in each turn. This mechanism also increases the window in which the model understands changes in emotion, i.e., the emotion of each utterance is judged jointly by information from at least two utterances, and can increase the probability of correct classification when the sentiment state changes.

For the current utterance, we select the GRU network corresponding to the next speaker to update the listener's emotion state.

$$e_i^{S(u_{i+1})} = GRU_s(e_{i-1}^{S(u_{i+1})}, CF_i^l) \tag{5}$$

The listener here is the chivalrous listener, precisely defined as the speaker of the next time step, expressed as the hearer for ease of understanding. Since the listener knowledge paragraph uses the next utterance as a fact, to ensure the validity of the commonsense knowledge paragraph, when a dialogue consists of more than two participants, the listener's commonsense knowledge feature is only used for updating the status of the speaker of the next utterance, and the status of the other non-current speakers remains unchanged.

$$e_i^{p_j} = e_{i-1}^{p_j}, p_j \neq S(u_{i+1}) \wedge j \in \{1, 2, ..., M\} \tag{6}$$

Ultimately, we add the Scene Emotion Vector and the Speaker Emotion Vector of the same utterance together and input the result into the softmax layer after going through a linear layer to obtain the final emotion classification of each utterance.

$$\begin{gathered} p_i = soft\max(W_{s\max}(E_i^{scene} + E_i^{\text{speaker}}) + b_{s\max}) \\ \widehat{y_i} = \arg\max_k (p_i[k]) \end{gathered} \tag{7}$$

4. Results and Discussion

4.1. Datasets

Our experiments were conducted on two benchmark datasets for conversation emotion recognition tasks: **MELD** [40] and **IEMOCAP** [41].These datasets are widely recognized in the community for their diversity in emotional expression, multi-speaker dialogue structures, and rich multimodal annotations (e.g., text, audio, and video), making them well-suited for evaluating models' ability to understand emotional dynamics in realistic conversational settings.

IEMOCAP consists of dialogue videos performed by ten actors in pairs, making up five sessions in total, including both audio and textual modalities. It comprises 7433 utterances across 151 dialogues, with each utterance labeled with an emotion. The emotions are categorized into six classes: happiness, sadness, neutral, anger, excitement, and frustration. Since IEMOCAP does not have a predefined split for training/validation/testing, to ensure fairness in subsequent comparisons, we follow previous studies [20] by training on the first four sessions and testing on the last session.

MELD features multi-person dialogue videos from the TV show "Friends", also including corresponding audio and video modalities. It includes 1433 dialogues and 13,708 utterances, with each utterance assigned an emotion label. The labels categorize the emotions into seven classes: anger, disgust, sadness, joy, surprise, fear, or neutral. We adhere to the predefined training/validation/testing split in MELD to maintain the fairness of our experimental results.

The dataset partitioning is summarized in Table 2, and the corresponding label distribution is illustrated in Figure 4.

Table 2. Statistics of datasets.

Dataset	Dialogue			Utterances		
	Train	Valid	Test	Train	Valid	Test
IEMOCAP	108	12	31	5163	647	1623
MELD	1038	114	280	9989	1109	2610

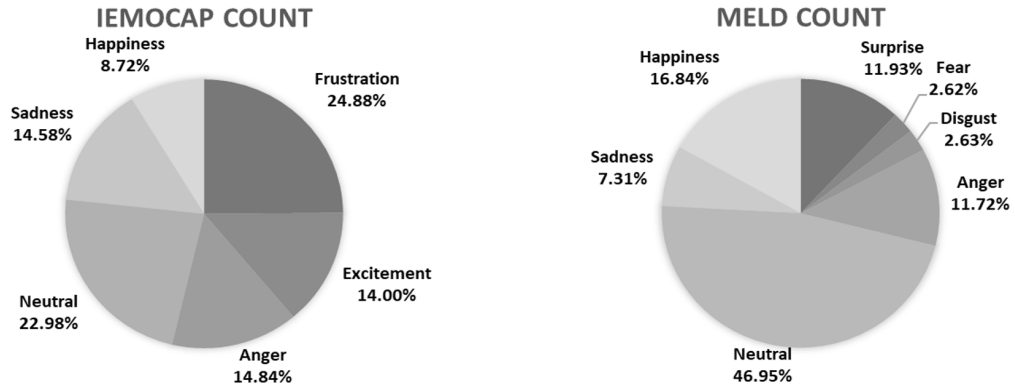

Figure 4. The class distribution of IEMOCAP and MELD.

4.2. Training Setup

We leverage both audio, visual and textual modalities in the MELD and IEMOCAP datasets. The specific dimensionality of different modalities is shown in Table 3.

Table 3. Feature dimensionality for different modalities.

Modality	MELD	IEMOCAP
Textual	600	100
Audio	300	100
Visual	300	100
Commonsense Knowledge	600	100

The common settings for both datasets include 30 training epochs, a batch size of 64, the Adam optimizer [42], L2 regularization weight λ of 0.001, and a dropout rate of 0.2. For the MELD dataset, the learning rate is set to 0.0001, and for the IEMOCAP dataset, it is set to 0.0003.

Both the IEMOCAP and MELD datasets exhibit severe class imbalance issues, as shown in Figure 4.

To address class imbalance, we employ Focal Loss [43] during training. By introducing a focal factor γ, Focal Loss directs the model to focus on difficult, misclassified examples and prevents over-representation of majority classes in the loss function. For both datasets, γ is set to 2.0; since the balancing parameter α only applies to modifying binary classification loss weights, we use the transformed class weights $\hat{weight}$ as a substitute.

The initial class weight for class i is computed as:

$$\text{weight}_i = \frac{1}{\log(\text{freq}_i + \epsilon)} \tag{8}$$

To stabilize training, we apply logarithmic compression and linear scaling to the class weights, which suppresses extreme values while preserving minority class contributions:

$$\begin{aligned} \text{weight}' &= \frac{\log(1 + \text{weight})}{\max(\log(1 + \text{weight}))} \\ \hat{weight} &= \frac{1}{2} \times weight' + \frac{1}{2} \end{aligned} \tag{9}$$

4.3. Evaluation Metrics

Similarly, due to the class imbalance in the datasets, we adopt weighted Average accuracy (WA-Acc) and weighted Average F1-score (WA-F1) as the evaluation metrics for overall model performance. Additionally, for a more detailed assessment and analysis, we compute the F1-score for each individual class separately.

The formula for computing the F1-score for a single class$i$ is:

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{10}$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{11}$$

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{12}$$

The formula for computing accuracy (ACC) for a single class$i$ is as follows:

$$\text{Acc}_i = \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \tag{13}$$

Among them, $\text{TP}_i, \text{TN}_i, \text{FP}_i, \text{FN}_i$ represent the True Positives, True Negatives, False Positives, and False Negatives of the predicted class, respectively.

The formulas for weighted average accuracy and weighted average F1-score are as followed:

$$\text{WA-Acc} = \sum_{i=1}^{N} w_i \times \text{Acc}_i \tag{14}$$

$$\text{WA-F1} = \sum_{i=1}^{N} w_i \times F1_i \tag{15}$$

4.4. Comparison with Other SOTA Methods

4.4.1. SSEAN-Uni

There are currently few attempts to enhance multimodal models with commonsense knowledge, and these approaches lack representativeness. Therefore, to ensure a fair evaluation while verifying the effectiveness of our proposed model, particularly the three-level commonsense knowledge dynamic filtering module, we first compare our model (SSEAN-Uni, with its structure shown in Figure 5) with other state-of-the-art (SOTA) models using only the text modality enhanced with commonsense knowledge. The results are presented in Table 4.

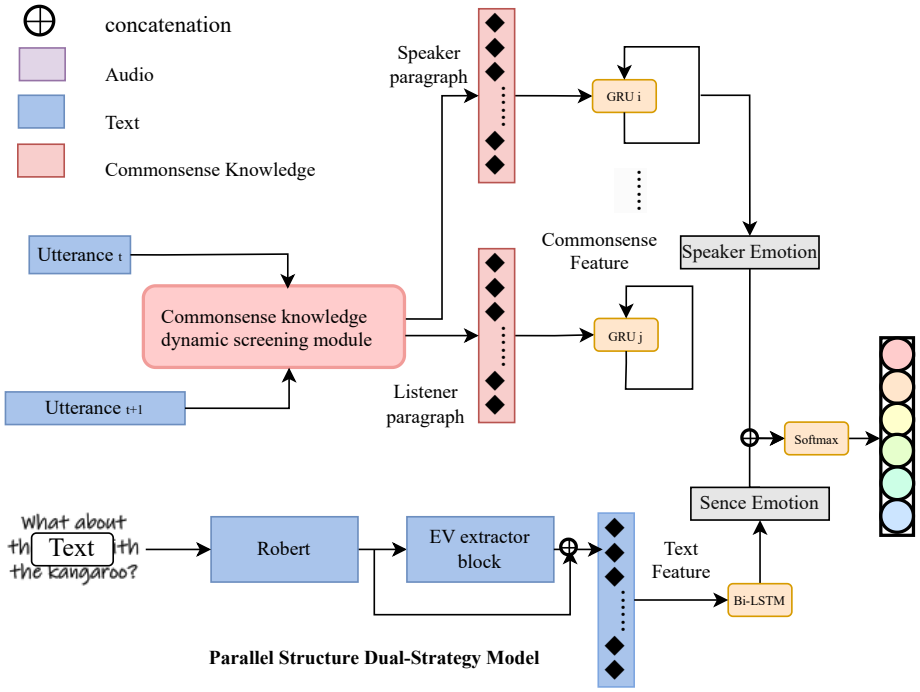

Figure 5. Framework illustration of commonsense knowledge-Enhanced Scene-Speaker Awareness Model for ERC.

To verify the stability and robustness of the model, we trained it using five different seeds and computed the mean, standard deviation, and Coefficient of Variation (CV) of the results, shown in Table 5. A CV value of less than 1 indicates that the model's results are stable.

SSEAN-Uni achieved superior F1 performance compared to all state-of-the-art (SOTA) models on the IEMOCAP dataset, attaining a score of 72.12, which achieves the best performance among all baseline models. On the MELD dataset, SSEAN-Uni also outperformed most models, ranking just slightly behind EmotionIC and InstrucERC. The performance gap with InstrucERC is minimal, and through seed adjustments, SSEAN-Uni can achieve results (66.38%) exceeding EmotionIC. However, compared to InstrucERC, there remains a

more significant performance gap. Since InstrucERC is a generative multi-task framework based on large language models (LLMs), we hypothesize that the complexity of emotional states in the MELD dataset poses a challenge for SSEAN-Uni, given its model size and the length constraints of single inputs. This limitation may hinder its ability to fully capture long-term emotional dependencies. Nevertheless, SSEAN-Uni offers a substantial advantage in terms of computational efficiency.

Table 4. The overall F1 scores of ERC on the IEMOCAP and MELD datasets.

Models	IEMOCAP WA-F1	MELD WA-F1
COSMIC [20]	65.28	65.21
CauAIN [44]	67.61	65.46
TodKat [21]	61.33	65.47
SKAIG [45]	66.96	65.18
EmotionIC [46]	69.44	66.32
MKFM [47]	68.88	65.66
InstrucERC [48]	71.39	**69.27**
SSEAN-Uni	**72.12**	66.17

Bold values indicate the best performance on the corresponding dataset.

Table 5. Experimental results with different random seeds.

SEED	IEMOCAP	MELD
0	72.44	66.03
42	71.88	65.97
100	72.21	66.12
1000	72.03	66.38
4027	72.12	66.17
mean	72.13	66.13
standard deviation	0.209	0.158
Coefficient of Variation	0.289%	0.239%

4.4.2. SSEAN-Multi

In Tables 6 and 7, we showcase the performance comparison of the SSEAN-Multi model with other state-of-the-art (SOTA) multimodal models on the emotion recognition in conversation (ERC) task on the IEMOCAP and MELD datasets. The experiment results demonstrate that SSEAN achieves the latest performance benchmarks on both datasets.

Table 6. The F1 results of ERC on the IEMOCAP dataset.

Models	Happy	Sad	Neutral	Angry	Excited	Frustrated	WA-F1	WA-Acc
DialogueRNN [9]	32.2	80.26	57.89	62.82	73.87	59.76	62.89	63.52
Emocaps [10]	71.91	85.06	64.48	**68.99**	78.41	66.78	71.77	-
DialogueGCN [14]	42.75	84.54	63.54	64.19	63.08	66.99	64.108	65.25
MMGCN [49]	51.57	80.48	57.69	53.95	72.81	57.33	62.89	63.22
MM-DFN [16]	42.22	78.98	66.42	69.77	75.56	66.33	68.18	68.21
BiF-BiAGRU [50]	54.50	72.70	59.40	61.00	66.60	61.60	63.00	62.80
SSEAN-Multi	**73.72**	**87.10**	**69.09**	68.39	**79.78**	**68.54**	**73.94**	**73.91**

Bold values indicate the best performance on the corresponding dataset.

Table 7. The F1 results of ERC on the MELD dataset.

Models	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry	WA-F1	WA-Acc
DialogueRNN [9]	76.97	47.69	-	20.41	50.92	-	45.52	57.66	60.31
Emocaps [10]	77.12	**63.19**	3.03	**42.52**	57.50	7.69	**57.54**	64.00	-
DialogueGCN [14]	72.10	41.70	2.80	21.80	44.20	6.70	36.50	54.70	54.90
MMGCN [49]	76.97	47.69	-	20.41	50.92	-	45.52	57.66	60.31
MM-DFN [16]	77.76	50.69	8.00	38.50	54.70	11.80	43.50	60.80	60.80
UniF-BiAGRU [50]	76.40	49.70	11.50	27.00	52.40	14.00	39.40	58.10	60.30
SSEAN-Multi	**80.02**	58.80	**27.27**	41.40	**64.40**	**36.07**	52.28	66.43	**67.04**

Bold values indicate the best performance on the corresponding dataset.

On the IEMOCAP dataset, the SSEAN model significantly outperforms all other models, obtaining the highest Weighted Average F1 Score (WA-F1) of 73.95% and an accuracy of 73.96%. This performance is notably robust across all categories except "Angry", indicating the model's robustness in recognizing complex emotional states. Particularly, the SSEAN model shows a significant improvement in the "Happy" category, with notable improvements in "Sad" and "Excited" categories as well, highlighting SSEAN's capability to discern subtle emotional expressions. This outstanding performance can be attributed to SSEAN's comprehensive contextual and speaker-level modeling, enabling it to more accurately capture emotional dynamics.

In the MELD dataset, SSEAN once again sets a new benchmark, with a WA-F1 score of 66.43% and an accuracy of 67.04%. It exhibits significant improvements across all categories. Notably, SSEAN's effective utilization of multimodal information and dynamic filtering mechanisms to extract and leverage relevant emotional cues is evidenced by its performance in the "Fear" and "Disgust" categories, which are traditionally challenging to model due to their subtle expressions and dependence on the speaker's state information.

SSEAN's performance enhancement stems from the effective utilization of a vast amount of information. The overall leading results affirm the efficacy of our model structure in handling information from multiple sources and modalities and reducing the accumulation of redundant information.

4.5. Ablation Study and Analysis

4.5.1. SSEAN-Uni

To further validate our model's proficient performance in processing commonsense knowledge and effectively avoiding the introduction of redundant information, we conducted extensive ablation experiments.

The ablation studies shown in Table 8 assessed the impact of components—such as the audio modality, commonsense knowledge, and speaker/scene emotion vectors—on the performance of SSEAN on the IEMOCAP and MELD datasets, highlighting the importance of each component for achieving state-of-the-art emotion recognition in conversation (ERC).

Notably, the most dramatic performance drop is observed when the Global Emotion vector is removed, with WA-F1 scores plummeting by 36.72% and 12.93% on IEMOCAP and MELD, respectively. This decline accentuates the critical role of the fundamental vector in capturing the core emotional tendencies across conversations, underlining its importance in the model's framework. This finding aligns with conclusions drawn from many other studies.

Table 8. The F1 results of Ablation Study.

	IEMOCAP		MELD	
	WA-F1	**Accuracy**	**WA-F1**	**Accuracy**
w/o CK dynamic screening module	70.69 (↓ 1.43)	70.57 (↓ 1.66)	62.04 (↓ 4.13)	63.35 (↓ 3.07)
w/o speaker identity modeling	71.48 (↓ 0.64)	71.26 (↓ 0.97)	64.69 (↓ 1.48)	65.41 (↓ 1.01)
w/o Speaker Emotion vector	69.82 (↓ 2.3)	69.53 (↓ 2.7)	63.22 (↓ 2.95)	64.03 (↓ 2.39)
w/o Global Emotion vector	35.40 (↓ 36.72)	37.79 (↓ 34.44)	53.24 (↓ 12.93)	56.82 (↓ 9.6)
SSEAN	**72.12**	**72.23**	**66.12**	**66.56**

Bold values indicate the best performance on the corresponding dataset. The ↓ indicate performance decrease compared to the full model.

Specifically, removing the CK dynamic screening module led to a 1.43% drop in WA-F1 on IEMOCAP and a 4.13% drop on MELD. This decrease is comparable to or even greater than the impact of not using commonsense knowledge, emphasizing the three-level dynamic screening module's crucial role in filtering relevant knowledge and reducing noise in the commonsense knowledge base, thus, ensuring the quality of knowledge integration.

Excluding speaker identity modeling also resulted in significant performance drops, particularly on the multi-party MELD dataset (1.48%). This underscores the necessity of capturing speaker-specific emotional states and interactions to avoid the introduction of redundant or irrelevant information due to speaker changes in multi-turn dialogues.

4.5.2. SSEAN-Muti

We investigated the impact of integrating commonsense knowledge with different modality combinations on model performance, as shown in Table 9. As expected, the tri-modal configuration achieved the best performance compared to bi-modal settings.

For the single-text modality model enhanced with commonsense knowledge, on the IEMOCAP dataset, adding any additional modality consistently improved the model's WA-F1 and WA-Acc. Similarly, on the MELD dataset, incorporating any new modality led to improvements in WA-Acc. This demonstrates that our Dual-Strategy Framework effectively utilizes multimodal information while mitigating the interference caused by redundant information across different modalities.

Table 9. Performance on IEMOCAP and MELD datasets.

Methods	IEMOCAP		MELD	
	WA-F1	**WA-Acc**	**WA-F1**	**WA-Acc**
Text (w CF)	72.12	72.23	66.17	66.42
Text (w CF) + Visual	73.36 ↑	73.32 ↑	65.68	65.54
Text (w CF) + Audio	73.14 ↑	73.19 ↑	65.85	66.73 ↑
Text (w CF) + Audio + Visual	**73.94** ↑	**73.91** ↑	**66.43** ↑	**67.04** ↑

Bold values indicate the best performance on the corresponding dataset. The ↑ indicates performance improvement compared to the text-only modality.

In the MELD dataset, although the model's overall performance slightly declined when only the textual and visual modalities were incorporated, a closer analysis of individual emotion categories (shown in Appendix B) reveals that the visual modality contributes to higher F1 scores in categories such as Neutral, Joy, and Angry.

4.6. Case Study

Commonsense knowledge occasionally incorporates extraneous information that may not align with the actual facts, thereby impeding the training process of models. However,

dynamic filtering mechanisms have the capability to mitigate this impact by selectively filtering out such irrelevant information.

Through existing research, we have identified the superior ability of commonsense knowledge in facilitating the transition of emotions in conversations and the classification of similar emotional categories. However, this introduces some challenges. Figure 6 presents a study case from the MELD dataset, where the conversation is entirely neutral. For traditional models that classify emotions solely based on text, this does not pose much of a challenge. However, models incorporating commonsense knowledge often make errors in such scenarios. As shown in Figure 6, the commonsense knowledge generated from the third utterance includes a listener's reaction that does not match reality. This kind of commonsense knowledge introduces sorrowful emotional information into subsequent judgments, leading to misclassification of the utterance's emotion. Through the dynamic filtering module for commonsense knowledge, such unrealistic commonsense information is filtered out, thus, removing sorrowful emotional information and avoiding classification errors.

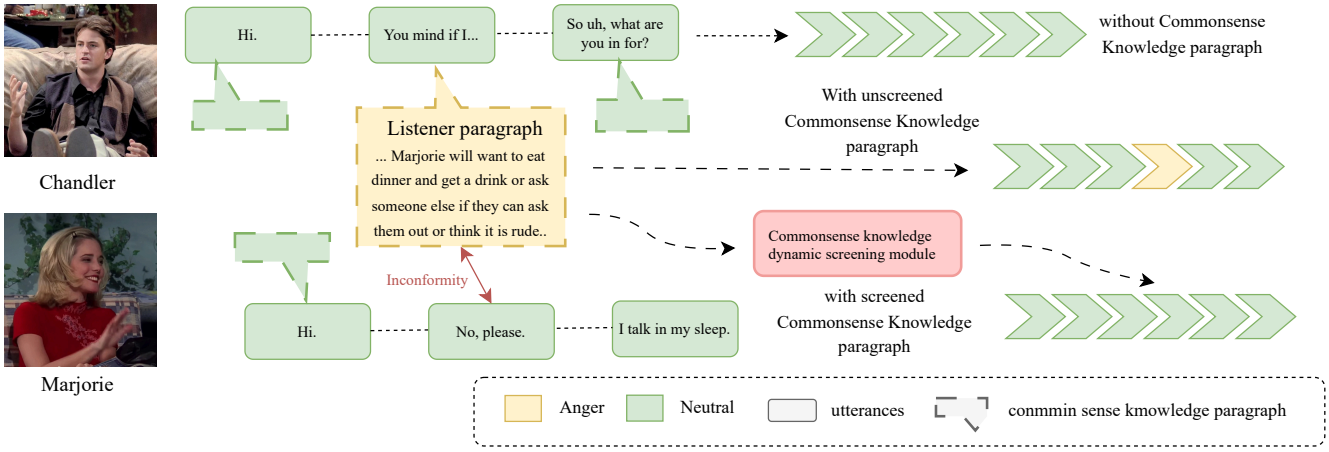

Figure 6. Case study from MELD dataset.

5. Conclusions

In this paper, we proposed the Scene-Speaker Emotion Aware Network (SSEAN) to address key challenges in emotion recognition in conversation (ERC). We introduced a dual-strategy framework that effectively models both global conversational context and speaker-specific emotional dynamics by leveraging two parallel recurrent networks. Our model captures long-term dependencies in dialogue while mitigating the interference of speaker identity information on emotional representation. Furthermore, we designed a three-level dynamic filtering module to refine and enhance the utilization of commonsense knowledge, improving its reliability and effectiveness in ERC tasks.

Our experiment results on the IEMOCAP and MELD datasets demonstrate that SSEAN achieves state-of-the-art (SOTA) performance in both single-text modality and multimodal settings. Further analysis confirms the effectiveness of multimodal integration and commonsense knowledge enhancement in ERC. Additionally, our commonsense knowledge filtering strategy significantly reduced noise from unreliable external knowledge, allowing SSEAN to make more accurate emotion predictions specifically in those classes with few samples. The model's stability and robustness were further validated through multiple training runs with different random seeds, showing low variance in performance, which underscores the statistical significance of our results.

While our method improves the quality and relevance of integrated commonsense knowledge, it remains limited in modeling long-range knowledge that requires reasoning across

multiple dialogue turns. Future work may involve leveraging large language models (LLMs) to enable extended context understanding and more advanced commonsense reasoning.

Author Contributions: Conceptualization, B.L. and C.L.; methodology, B.L.; software, B.L.; validation, H.Z., L.L. and B.L.; formal analysis, B.L.; investigation, Y.G.; resources, Y.G.; data curation, B.L.; writing—original draft preparation, B.L.; writing—review and editing, Y.G., H.L. and S.W.; visualization, B.L.; supervision, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (grant numbers 2021ZD0110400 and 2021ZD0110404), the National Natural Science Foundation of China (grant numbers 62271377 and 62201407), and the Key Research and Development Program of Shaanxi Province (grant numbers 2021ZDLGY0106, 2022ZDLGY01-12, 2023YBGY244, 2023QCYLL28, 2024GX-ZDCYL-02-08, and 2024GX-ZDCYL-02-17). The APC was funded by the authors.

Data Availability Statement: The original data presented in the study are openly available in https://sail.usc.edu/iemocap/ (accessed on 24 June 2025) and https://github.com/declare-lab/MELD (accessed on 24 June 2025).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Comparative Analysis of Knowledge Representation Filtering

By adding a two-layer feedforward classification head to the knowledge representations, we directly applied the commonsense knowledge to emotion classification. As shown in Table A2, the filtered knowledge consistently leads to improved classification accuracy across all emotion categories. The results of ten-fold cross-validation further confirm that the filtered knowledge achieves consistently better overall performance, providing additional evidence for the effectiveness of our proposed knowledge filtering module.

Table A1. F1-score comparison of knowledge representation before and after filtering.

Emotion Category	Pre-Filtering F1	Post-Filtering F1	Improvement
0 (Neutral)	0.8966	0.9120	+0.0154
1 (Surprise)	0.2444	0.4286	+0.1842
2 (Fear)	0.1600	0.2500	+0.0900
3 (Sadness)	0.3077	0.4828	+0.1751
4 (Joy)	0.1527	0.3089	+0.1562
5 (Disgust)	0.2500	0.5000	+0.2500
6 (Anger)	0.1644	0.4235	+0.2591
Accuracy	0.8125	0.8462	+3.37%
Weighted Average F1	0.7595	0.8086	+4.91%
Macro Average F1	0.3108	0.4723	+16.14%

Appendix B. Multimodal Fusion Analysis on MELD

In the MELD dataset, although incorporating only the textual and visual modalities leads to a slight decline in overall model performance, a more fine-grained analysis of individual emotion categories indicates that the visual modality contributes to improved F1 scores in categories such as Neutral, Joy, and Angry. Compared to other emotion types, Joy and Angry are more strongly associated with pronounced bodily movements and facial expressions—such as smiling, frowning, or dynamic gestures—which are effectively captured by visual features. This finding suggests that the video features extracted by the

3D-CNN still offer complementary information, especially for recognizing emotions that rely on non-verbal cues.

Table A2. MELD: Text (w/ CF) + Visual and Text (w/ CF) configurations.

Emotion Label	Text (w/ CF) + Visual	Text (w/ CF)
Neutral	0.7937 ↑	0.7892
Surprise	0.5787	0.5921
Fear	0.1707	0.3077
Sadness	0.3701	0.4377
Joy	0.6551 ↑	0.6462
Disgust	0.3248	0.3193
Anger	0.5330 ↑	0.5262
WA-F1	65.68	66.17
WA-Acc	66.54	66.42

The ↑ indicates performance improvement compared to the text-only modality.

Appendix C. Evaluation of the Inference Efficiency

Table A3. Evaluation of the inference efficiency of the SSEAN-Multi model on two datasets.

Dataset	Total Parameters	Inference Time (ms/Sample)	FLOPs (Forward Pass)
IEMOCAP	7.3 M	19.03	1.2 GFLOPs
MELD	23.7 M	21.30	1.5 GFLOPs

Despite a larger number of parameters, the model maintains acceptable inference time and computational cost.

It is worth noting that the parameter count on the MELD dataset is relatively high, primarily due to its dialogues often involving multiple speakers—up to nine in extreme cases—which is rare in everyday conversations.

To address this scalability challenge, our model adopts a dynamic GRU instantiation strategy, where GRU units are assigned only to speakers who actually appear in a given dialogue. This design effectively avoids redundant computation in scenarios with fewer participants. In practice, only two to three GRU units are typically activated during inference, which significantly reduces computational overhead. This observation is further supported by the reported FLOPs and inference time.

Additionally, when speaker identity information is unavailable or ambiguous in the data, the model naturally degrades to a simplified two-GRU version, maintaining scalability and adaptability. This design choice is further validated by our FLOP analysis, which supports the efficiency of the proposed framework.

In comparison, although the SSEAN-Multi model contains more parameters than its SSEAN-Uni counterpart, the latter achieves better inference efficiency, making it a more favorable choice when runtime performance is a priority.

Appendix D. Error Analysis Based on the Confusion Matrix of SSEAN-Multi

To better understand where our model underperforms, we conduct an in-depth analysis based on the confusion matrix of the IEMOCAP and MELD dataset.

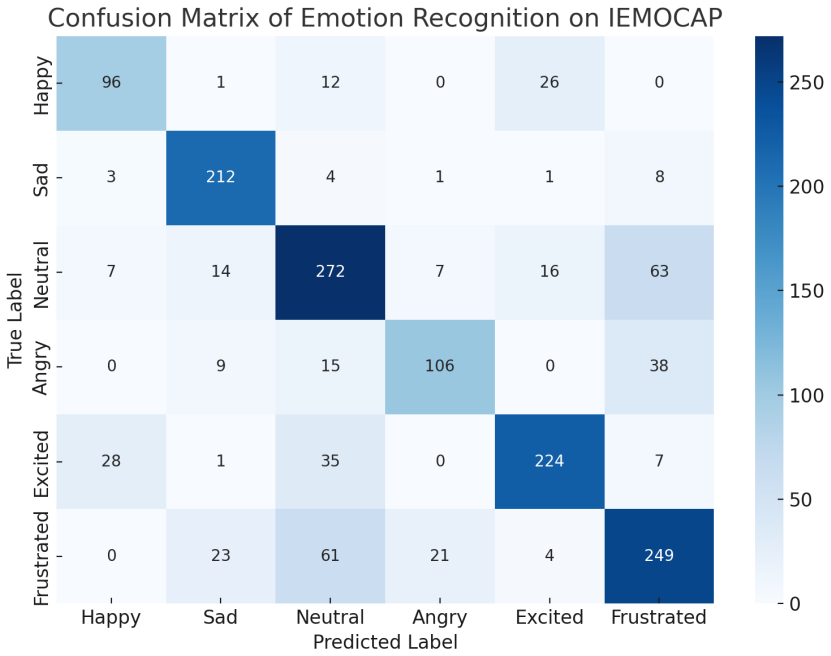

Figure A1. Confusion matrix of emotion recognition on IEMOCAP.

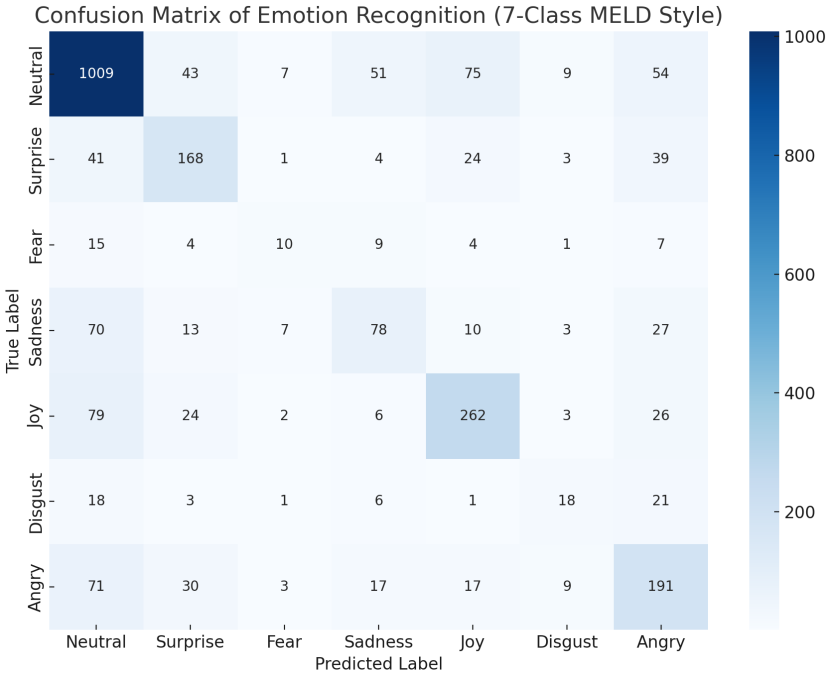

Figure A2. Confusion matrix of emotion recognition on MELD.

From the confusion matrices of both the IEMOCAP and MELD datasets, consistent misclassification patterns can be observed, revealing the model's challenges in handling ambiguous emotional boundaries and low-resource emotion categories.

In the IEMOCAP dataset, although Frustrated and Excited exhibit some misclassifications, SSEAN still achieves the highest classification accuracy on these two categories. Notably, most of their misclassified instances fall into emotionally similar classes: Frustrated is often confused with other negative emotions such as Neutral, Sad, or Angry, while Excited tends to be misclassified as positive or neutral emotions, such as Happy or Neutral. This reflects a tendency toward semantic polarity-consistent misclassification, which is

generally more tolerable in practical applications and suggests that the model captures emotional polarity effectively, even if fine-grained distinctions are not always accurate.

In the MELD dataset, the classification of Fear and Disgust proves significantly more challenging. These categories are frequently misclassified as Sadness, Neutral, or Angry. The main causes are twofold: first, these categories suffer from extreme data scarcity, which limits the model's ability to learn distinctive features; second, their linguistic and acoustic similarity to other negative emotions contributes to confusion, and the external commonsense knowledge introduced by the model is not yet sufficiently effective in modeling the emotional causality associated with Fear and Disgust.

Moreover, both datasets exhibit a notable tendency for Neutral to absorb a large number of misclassifications. In cases of emotional ambiguity, incomplete multimodal input, or unclear signals, the model tends to make conservative predictions toward Neutral. This is closely related to the overrepresentation of Neutral samples in both datasets and suggests that the model adopts a risk-averse strategy when facing uncertainty, favoring safety over discrimination.

References

1. Khare, S.K.; Blanes-Vidal, V.; Nadimi, E.S.; Acharya, U.R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* **2023**, *102*, 102019. [CrossRef]
2. Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* **2019**, *7*, 100943–100953. [CrossRef]
3. Baker, B.; Mills, K.A.; McDonald, P.; Wang, L. AI, concepts of intelligence, and chatbots: The "Figure of Man," the rise of emotion, and future visions of education. *Teach. Coll. Rec.* **2023**, *125*, 60–84. [CrossRef]
4. Olujimi, P.A.; Ade-Ibijola, A. NLP techniques for automating responses to customer queries: A systematic review. *Discov. Artif. Intell.* **2023**, *3*, 20. [CrossRef]
5. Sadhu, A.K.R.; Parfenov, M.; Saripov, D.; Muravev, M.; Sadhu, A.K.R. Enhancing Customer Service Automation and User Satisfaction: An Exploration of AI-powered Chatbot Implementation within Customer Relationship Management Systems. *J. Comput. Intell. Robot.* **2024**, *4*, 103–123.
6. Rodríguez-Ibánez, M.; Casánez-Ventura, A.; Castejón-Mateos, F.; Cuenca-Jiménez, P.M. A review on sentiment analysis from social media platforms. *Expert Syst. Appl.* **2023**, *223*, 119862. [CrossRef]
7. Tanna, D.; Dudhane, M.; Sardar, A.; Deshpande, K.; Deshmukh, N. Sentiment analysis on social media for emotion classification. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 911–915.
8. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
9. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6818–6825.
10. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion capsule based model for conversational emotion recognition. *arXiv* **2022**, arXiv:2203.13504.
11. Li, J.; Liu, Y.; Wang, X.; Zeng, Z. CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition. *arXiv* **2023**, arXiv:2307.15432. [CrossRef]
12. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: A comprehensive review. *Comput. Soc. Netw.* **2019**, *6*, 1–23. [CrossRef] [PubMed]
13. Geng, T.; Wu, C.; Zhang, Y.; Tan, C.; Xie, C.; You, H.; Herbordt, M.; Lin, Y.; Li, A. I-GCN: A graph convolutional network accelerator with runtime locality enhancement through islandization. In Proceedings of the MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture, Virtual, 18–22 October 2021; pp. 1051–1063.
14. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. 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), Hong Kong, China, 3–7 November 2019; pp. 154–164.

15. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.; Modi, A. 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, Seattle, WA, USA, 10–15 July 2022; pp. 4148–4164.
16. Hu, D.; Hou, X.; Wei, L.; Jiang, L.; Mo, Y. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7037–7041.
17. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed Acyclic Graph Network for Conversational Emotion Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual, 1 August 2021; Volume 1: Long Papers, pp. 1551–1560.
18. Yeomans, M.; Boland, F.K.; Collins, H.K.; Abi-Esber, N.; Brooks, A.W. A practical guide to conversation research: How to study what people say to each other. *Adv. Methods Pract. Psychol. Sci.* **2023**, *6*, 25152459231183919. [CrossRef]
19. Wei, W.L.; Wu, C.H.; Lin, J.C.; Li, H. Exploiting psychological factors for interaction style recognition in spoken conversation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 659–671. [CrossRef]
20. Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; Poria, S. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Virtual, 16–20 November 2020; pp. 2470–2481.
21. Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; He, Y. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual, 1 August 2021; Volume 1: Long Papers, pp. 1571–1582.
22. Speer, R.; Chin, J.; Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
23. Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N.A.; Choi, Y. Atomic: An atlas of machine commonsense for if-then reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, WI, USA, 27 January–1 February 2019; Volume 33, pp. 3027–3035.
24. Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; Choi, Y. *COMET: Commonsense Transformers for Knowledge Graph Construction*; Association for Computational Linguistics (ACL): Kerrville, TX, USA, 2019.
25. Fu, Y.; Wu, J.; Wang, Z.; Zhang, M.; Shan, L.; Wu, Y.; Liu, B. LaERC-S: Improving LLM-based Emotion Recognition in Conversation with Speaker Characteristics. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 6748–6761.
26. Liu, H.; Wei, R.; Tu, G.; Lin, J.; Jiang, D.; Cambria, E. Knowing What and Why: Causal emotion entailment for emotion recognition in conversations. *Expert Syst. Appl.* **2025**, *274*, 126924. [CrossRef]
27. Shen, X.; Huang, X.; Zou, S.; Gan, X. Multimodal knowledge-enhanced interactive network with mixed contrastive learning for emotion recognition in conversation. *Neurocomputing* **2024**, *582*, 127550. [CrossRef]
28. Zhang, X.; Wang, M.; Zhuang, X.; Zeng, X.; Li, Q. CDEA: Causality-Driven Dialogue Emotion Analysis via LLM. *Symmetry* **2025**, *17*, 489. [CrossRef]
29. Lu, W.; Hu, Z.; Lin, J.; Wang, L. LECM: A model leveraging emotion cause to improve real-time emotion recognition in conversations. *Knowl.-Based Syst.* **2025**, *309*, 112900. [CrossRef]
30. Kumar, S.; Akhtar, M.S.; Chakraborty, T. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. *arXiv* **2023**, arXiv:2310.13080.
31. Gou, Z.; Long, Y.; Sun, J.; Gao, K. TG-ERC: Utilizing three generation models to handle emotion recognition in conversation tasks. *Expert Syst. Appl.* **2025**, *268*, 126269. [CrossRef]
32. Tu, G.; Liang, B.; Mao, R.; Yang, M.; Xu, R. *Context or Knowledge Is Not Always Necessary: A Contrastive Learning Framework for Emotion Recognition in Conversations*; Association for Computational Linguistics ACL: Kerrville, TX, USA, 2023; pp. 14054–14067.
33. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
34. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [CrossRef]
35. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
36. Haddad, J.; Lézoray, O.; Hamel, P. 3d-cnn for facial emotion recognition in videos. In Proceedings of the Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, 5–7 October 2020; Proceedings, Part II 15; Springer: Berlin/Heidelberg, Germany, 2020; pp. 298–309.
37. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.
38. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

39. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
40. Kim, W.R.; Mannalithara, A.; Heimbach, J.K.; Kamath, P.S.; Asrani, S.K.; Biggins, S.W.; Wood, N.L.; Gentry, S.E.; Kwong, A.J. MELD 3.0: The model for end-stage liver disease updated for the modern era. *Gastroenterology* **2021**, *161*, 1887–1895. [CrossRef]
41. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
44. Zhao, W.; Zhao, Y.; Lu, X. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22). International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 23–29 July 2022; pp. 4524–4530. [CrossRef]
45. Li, J.; Lin, Z.; Fu, P.; Wang, W. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021*; Association for Computational Linguistics ACL: Kerrville, TX, USA, 2021; pp. 1204–1214. [CrossRef]
46. Liu, Y.; Li, J.; Wang, X.; Zeng, Z. EmotionIC: Emotional Inertia and Contagion-Driven Dependency Modeling for Emotion Recognition in Conversation. *Sci. China Inf. Sci.* **2024**, *67*, 182103. [CrossRef]
47. Tu, G.; Liang, B.; Qin, B.; Wong, K.F.; Xu, R. An Empirical Study on Multiple Knowledge from ChatGPT for Emotion Recognition in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023*; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 12160–12173. [CrossRef]
48. Lei, S.; Dong, G.; Wang, X.; Wang, K.; Qiao, R.; Wang, S. InstructERC: Reforming Emotion Recognition in Conversation with Multi-task Retrieval-Augmented Large Language Models. *arXiv* **2023**, arXiv:2309.11911.
49. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1437–1445.
50. Jiao, W.; Lyu, M.R.; King, I. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), New York, NY, USA, 7–12 February 2020; AAAI Press: Washington, DC, USA, 2020; pp. 7981–7988. [CrossRef]