Watch the Speakers: A Hybrid Continuous Attribution Network for Emotion Recognition in Conversation with Emotion Disentanglement

***Abstract*—Emotion Recognition in Conversation (ERC) has attracted widespread attention in the natural language processing field due to its enormous potential for practical applications. Existing ERC methods face challenges in achieving generalization to diverse scenarios due to insufficient modeling of context, ambiguous capture of dialogue relationships and overfitting in speaker modeling. In this work, we present a Hybrid Continuous Attributive Network (HCAN) to address these issues in the perspective of emotional continuation and emotional attribution. Specifically, HCAN adopts a hybrid recurrent and attention-based module to model global emotion continuity. Then a novel emotional attribution encoding (EAE) is proposed to model intra- and inter-emotional attribution for each utterance. Moreover, aiming to enhance the robustness of the model in speaker modeling and improve its performance in different scenarios, A comprehensive loss function emotional cognitive loss $\mathcal{L}_{\mathrm{EC}}$ is proposed to alleviate emotional drift and overcome the overfitting of the model to speaker modeling. Our model achieves state-of-the-art performance on three datasets, demonstrating the superiority of our work. Another extensive comparative experiments and ablation studies on three benchmarks are conducted to provided evidence to support the efficacy of each module. Further exploration of generalization ability experiments shows the plug-and-play nature of the EAE module in our method.**

***Index Terms*—natural language processing, emotion recognition in conversation, context modeling, dialogue relationship**

I. Introduction

Emotion Recognition in conversation (ERC) is a rapidly growing research field within natural language processing (NLP) that focuses on identifying the emotions conveyed in each utterance of a conversation. Different from the single sentence's emotional classification in explicit sentiment analysis [1]–[4], this task contains samples with vastly different conversation lengths, ambiguous emotional expressions, and complex conversational relationships. Fig. 1 illustrates an example of the conversation scenario, where the utterance to be predicted (the last utterance) is influenced by the historical utterances of that conversation. As expected, ERC task has attracted the attention of many researchers due to its potential applications in various fields such as political campaigning and public opinion analysis [5], [7], human-robot interaction [8] and empathic dialogue system [9], [10].

Pervious ERC methods generally formulate the task as a supervised learning task based on different architectures of neural networks. This places a significant demand on the model's ability to capture the context of each utterance and effectively utilize speaker information [11]. Moreover, various modeling methods for context and speaker have significantly

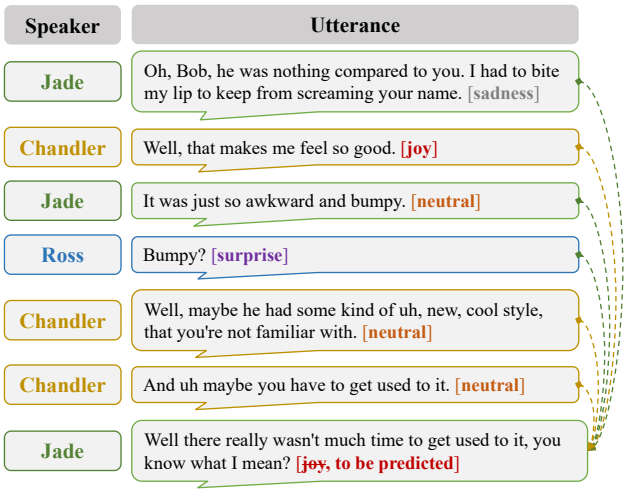

Fig. 1. A example for the conversation in the MELD dataset.

raised the baseline, but there are still two remaining challenges of ERC need to solve. (1) **Insufficient modeling of context.** Existing works on context modeling can be broadly categorized into two types: The recurrent based methods [12]–[15] focus on establishing more natural context temporal correlation. However, these methods may struggle to capture the global emotional continuity in long conversations. Although attention-based methods [16], [17], [19], [20] aim to aggregate emotional features at multiple levels, they may not be as effective as temporal models in capturing emotional continuity between speakers over time. These methods adopt a single and redundant network architecture, which results in a lack of generalization in context modeling. (2) **Ambiguous capture of dialogue relationships.** Studies [23], [24] provide evidence that generating emotional responses can effectively improve the performance of ERC models. It can be inferred that in real-life conversations, more direct conversational relationships often lead to more direct emotional transmission. Nonetheless, the ERC field still lacks of detailed modeling of the emotional influence within and between speakers in the perspective of dialogue relationship. (3) **Overfitting in speaker modeling.** In the ERC task, speakers often exhibit distinct characteristics in their emotional expressions due to differences in identity and personality. To better leverage this information, several studies have made significant contributions. Although intricate network designs have been developed from various perspectives,

such as speaker psychological states, dialogue memory, and relative positional relationships, these approaches have yielded limited results. Specifically, The models have encountered overfitting issues in different dialogue scenarios, which has hindered their effectiveness.

Therefore, these three limitations greatly hinder the application of ERC models in real-world scenarios, which is precisely what our work aims to address.

We have proposed HCAN to effectively address the aforementioned issues. To tackle the problem of insufficient context modeling, we propose Emotional Continuation Encoding (ECE) to extract more robust features in different conversation situations, which comprehensively utilizes both the recurrent units and the attention blocks. The *Attribution Theory* [21] proposes that a stimulus triggers perception, which leads individuals to consider the situation, and physiological reactions lead to cognitive interpretation of physiological changes, both of which together result in emotional expression. Drawing inspiration from the *Attribution Theory* and accurately capturing dialogue relationships, we present Emotional Attribution Encoding (EAE) based on IA-attention, which models the intra-attribution and inter-attribution of each sentence in an attribution perspective. Last but not least, emotional cognitive loss is introduced in ECE to effectively enhance the model's robustness and extend the applicability of the overall model. The Emotional Cognitive loss $\mathcal{L}_{EC}$ is composed of cross-entropy $\mathcal{L}_{cross}$, KL divergence $\mathcal{L}_{\mathrm{KL}}$ for predicting and recognizing emotions, and Adversarial Emotion Disentanglement loss $\mathcal{L}_{adv}$. Among them, cross-entropy calculation serves as the main emotional loss, KL divergence can alleviate emotional drift, and Adversarial Emotion Disentanglement loss can mitigate the overfitting of the model to speaker modeling.

Our contributions are three-fold:

(1) By combining the recurrent and attention-based approaches, our proposed ECE module achieves strong robustness in global emotion continuity modeling across different datasets, particularly demonstrating outstanding performance on long conversation samples.

(2) Consider capturing dialogue relationships in the perpective of *Attribution Theory*, we propose an original IA-attention to extract intra-attribution and inter-attribution features, which offers a more direct and accurate modeling of human emotional comprehension.

(3) Our model achieves state-of-the-art performance on three datasets, demonstrating the superiority of our work. The proposed EAE module is a plugin module that exhibits strong generalization and effectiveness across different baselines.

II. Related Work

A. Emotion Recognition of Conversation

The significant advancement of deep learning has greatly promoted the improvement of baseline performance in ERC tasks. Recently, ERC models can be categorized into two types: recurrent-based methods and attention-based methods.

1) Recurrent-based Methods: Through the use of a sequential network structure, recurrent-based methods have the potential to offer a more precise and authentic representation of the emotional dynamics present in a conversation: DialogueRNN is the first to utilize a recurrent neural network for monitoring both speaker states and global states in conversations. COSMIC is a conversational model that integrates commonsense knowledge to enhance its performance. This model injects commonsense knowledge into Gated Recurrent Units to capture features related to the internal state, external state, and intent state. The performance of SKAIG is enhanced by integrating action information inferred from the preceding context and the intention suggested by the subsequent context. DialogueCRN is designed with multi-turn reasoning modules that extract and integrate emotional clues. These modules perform an iterative process of intuitive retrieval and conscious reasoning, which imitates the distinctive cognitive thinking of humans. With the goal of achieving a comprehensive understanding of the dialogue, CauAIN first retrieves and enhances causal clues in the dialogue through an external knowledge base. Then, it models intra- and inter-speaker interactions using GRUs.

2) Attention-based methods: To enable the extraction of emotional features at both coarse-grained and fine-grained levels, attention-based methods often employ a variety of encoders and decoders with different levels and structures. KET extracts concepts related to non-pause words in neutral discourse from a knowledge base and enhances the semantic representation of vectors using a dynamic context graph attention mechanism. Finally, a hierarchical self-attention mechanism is utilized to model the dialogue level. By leveraging four distinct attention mechanisms, DialogXL utilizes the language model layers of XLNet to encode multi-turn dialogues that are arranged in a sliding window. By regarding the internal structure of dialogue as a directed acyclic graph to encode utterances, DAG-ERC offers a more intuitive approach to modeling the information flow between the distant conversation background and the nearby context. TODKAT, as proposed in [26], presents a language model (LM) that is enhanced with topics through an additional layer specialized in detecting them. This model also incorporates commonsense statements obtained from a knowledge base based on the dialogue context.

B. Dialogue Relation Extraction

The task of relationship extraction (RE) aims to identify the relationships that exist between pairs of entities within a document. While in dialogue scenarios, the task of extracting dialogue relations becomes more challenging due to the ellipsis of expression, the fuzziness of semantic reference and the presence of long-distance context dependencies.

DialogRE [20] introduced the first human-annotated dataset for dialogue relationship extraction (DiaRE), which aims to capture the relationships between two arguments that arise in predictive conversations. Building upon this dataset, Chen [7] proposed a DiaRE method based on a graphical attention network that constructs meaningful graphs connecting

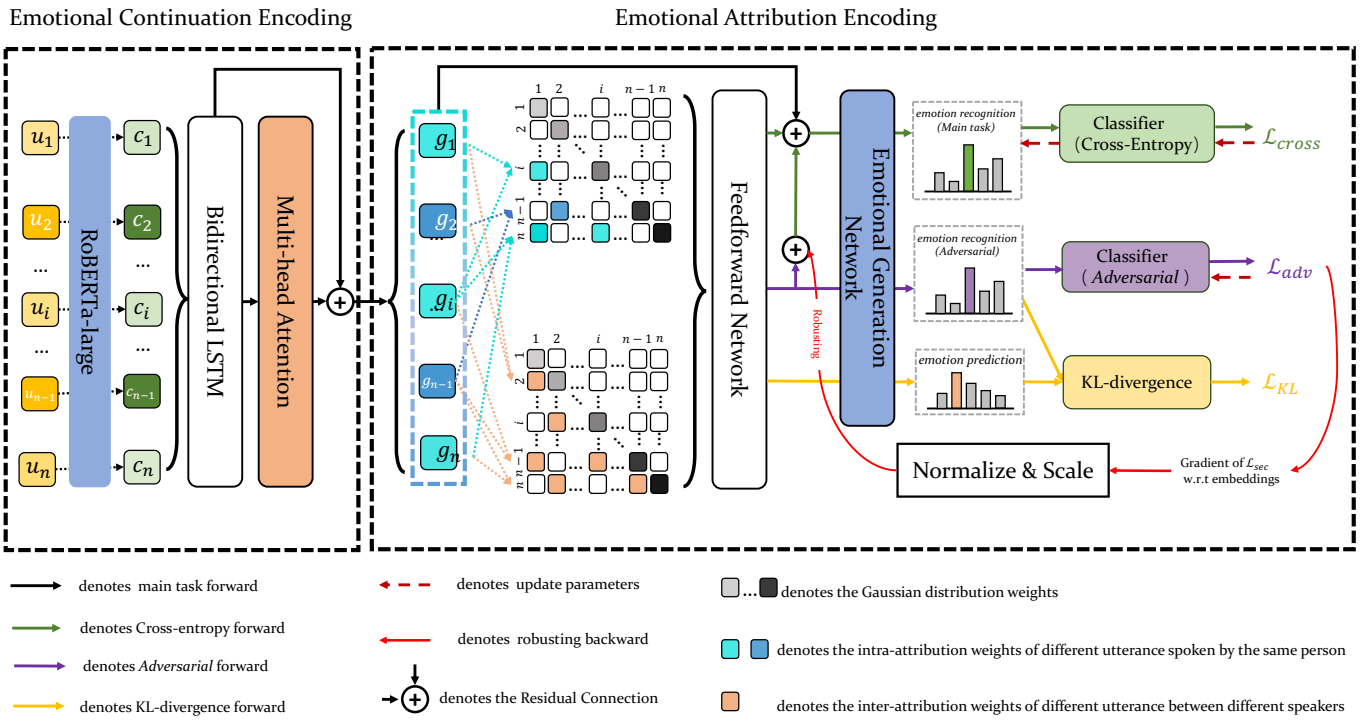

Fig. 2. The overall architecture of HCAN consisting of two main components, namely Emotional Continuation Encoding and Emotional Attribution Encoding.

speakers, entities, entity types, and corpus nodes to model the relationships between critical speakers. Similarly, Sun [8] proposed an utterance-aware graph neural network (ERMC-DisGCN) for ERMC, which leverages a relational convolution to propagate contextual information and takes into account the self-speaker dependency of interlocutors.

Despite the promising results achieved by the aforementioned methods, they have not been validated on the ERC dataset. Furthermore, unlike directly identifying the current emotional state based on DiaRE, our approach extracts dialogue relationships from an attributional perspective and adds an emotional prediction loss to the task, which better aligns with human thought processes and enhances the robustness of the model in different scenarios.

III. Methodology

In this section, we present the details of how to approach conversation modeling from a continuation-attribution perspective. The overview of HCAN is shown in Fig. 2, which is consist of Emotional Continuation Encoding and Emotional Attribution Encoding.

A. Task Statement

In the ERC task, the goal is to identify the emotion s_i of each utterance u_i in a conversation $[u_1, u_2, ..., u_N]$ by analyzing the dialogic context and the related speaker information p_i in speaker set $\{p_i, \ldots, p_M\}$, where the emotion should be selected from a pre-defined emotional target set S and each utterance corresponds to one speaker in the set of speakers.

B. Emotional Continuation Encoding

To mimic the natural conversational flow between speakers, the bidirectional LSTM is employed to encode the utterances' feature $\mathbf{c}_i \in \mathbb{R}^{d_u}$ in a temporal sequence as follows:

$$\mathbf{g}_i^l, \mathbf{h}_i = \overleftrightarrow{\text{LSTM}}(\mathbf{c}_i, \mathbf{h}_{i-1}) \tag{1}$$

where $\mathbf{h}_i \in \mathbb{R}^{2d_u}$ is the hidden state of the LSTM. Noted that the feature at the utterance-level of u_i is represented by $\mathbf{c}_i \in \mathbb{R}^{d_u}$, and it is obtained through the employment of the COSMIC method for extraction.

To avoid the vanishing of emotional continuity over long time spans, we utilized a multi-head attention module to aggregate the global information from the LSTM encoding result $\mathbf{G}^l$ as follows:

$$\mathbf{G} = \text{Multi-Attn}(W_Q\mathbf{G}^l, W_K\mathbf{G}^l, W_V\mathbf{G}^l) + \mathbf{G}^l \tag{2}$$

where $\mathbf{G}^l = [\mathbf{g}_1^l, \ldots, \mathbf{g}_n^l]$, $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_n]$, $\mathbf{g}_n^l \in \mathbb{R}^{2d_u}$, $\mathbf{g}_n \in \mathbb{R}^{2d_u}$ and W_Q, W_K, W_V are trainable parameters. The use of residual connections $+$ ensures that even in the worst-case scenario, the global emotional state degrades to a temporal emotional state, thereby enhancing the robustness of the model.

C. Emotional Attribution Encoding

Emotional Attribution Encoding is the core of this article, consisting of the IA-attention module and Emotional Cognitive loss. The IA-attention module efficiently captures the dialogue relationship and establishes emotional influence from the perspective of attribution. The Emotional Cognitive loss

effectively mitigates the overfitting of modeling on different datasets.

1) IA-attention: Inspired by the attribution theory of emotion, we examine the emotional influence in dialogue realtionships in an attributional prespective. Specially, we model emotional influence as intra-attribution and inter-attribution.

To achieve this, we introduce IA-attention, which is inspired by self-attention [22]. This method views each global utterance representation $\mathbf{g}_i$ as a query, which is mapped to intra-attribution partial space Q_a and inter-attribution partial space Q_e to get two different query embeddings $\mathbf{q}_{i_a}, \mathbf{q}_{i_e}$. Meanwhile, the historical utterance $[\mathbf{g}_1, \dots \mathbf{g}_{i-1}]$ are also projected to K and V partial space to obtain $\mathbf{k}_i$ and $\mathbf{v}_i$. To summarize, for each utterance, we apply different attribution attention matrices to get the intra-atribution weighted sum and inter-attribution weighted sum which are divided by each utterance's speaker p_i. The specific formula is as follows:

$$[\mathbf{q}_{i_a}; \mathbf{q}_{i_e}] = [W_{Q_a}; W_{Q_e}]\mathbf{g}_i \tag{3}$$

$$[\mathbf{k}_1, \dots, \mathbf{k}_n] = W_{K_{\mathrm{IA}}}[\mathbf{g}_1, \dots, \mathbf{g}_n] \tag{4}$$

$$[\mathbf{v}_1, \dots, \mathbf{v}_n] = W_{V_{\mathrm{IA}}}[\mathbf{g}_1, \dots, \mathbf{g}_n] \tag{5}$$

$$\tilde{\mathbf{v}}_i = \underbrace{\sum_{j<i, p_j=p_i} \frac{e^{\mathbf{q}_{i_a}^T \cdot \mathbf{k}_j}}{Z}\mathbf{v}_j}_{intra\text{-}attribution} + \underbrace{\sum_{j<i, p_j\neq p_i} \frac{e^{\mathbf{q}_{i_e}^T \cdot \mathbf{k}_j}}{Z}\mathbf{v}_j}_{inter\text{-}attribution} \tag{6}$$

where $W_{Q_a}, W_{Q_e}, W_{K_{\mathrm{IA}}}, W_{V_{\mathrm{IA}}} \in \mathbb{R}^{2d_u \times 4d_u}$ are trainable parameters, $\mathbf{q}_{i_a}, \mathbf{q}_{i_e}, \mathbf{k}_j, \mathbf{k}_i \in \mathbb{R}^{4d_u}$ and Z is the normalized factor.

To enable a more realistic perception in the dialogic relationship, the Gaussian Self-attention Mechanism [25] is introduced to distinguish the varying effects of dialogic temporal location. Assuming that the emotional attribution of historical utterances to the current utterance follows a normal distribution, the encoding results of the IA-attention module will be assigned weights that obey a Gaussian distribution, which is calculated as follows:

$$\hat{\mathbf{v}}_i = \sum_{j<i} \phi(d_{i,j}|\mu, \sigma)\tilde{\mathbf{v}}_j \tag{7}$$

where $\hat{\mathbf{v}}_i \in \mathbb{R}^{4d_u}$,$\phi$ is a Gaussian distribution, μ and σ are their corresponding learnable parameters, $d_{i,j}$ stands for distance measuring the turn-taking interval between speakers [25].

2) Emotional Cognitive Loss: The emotional overfitting of the ERC task mainly focuses on emotional drift and speaker modeling. Our proposed Emotional Cognitive loss $\mathcal{L}_{\mathrm{EC}}$ is mainly composed of basic cross-entropy $\mathcal{L}_{cross}$, KL divergence $\mathcal{L}_{\mathrm{KL}}$ for predicting and recognizing emotions, and Adversarial Emotion Disentanglement loss $\mathcal{L}_{adv}$. Among them, cross-entropy calculation is the main emotional loss, KL divergence can alleviate emotional drift, and Adversarial Emotion Disentanglement loss can overcome the overfitting of the model to speaker modeling.

Cross-entropy loss $\mathcal{L}_{cross}$, the key elements of which are computed as follows:

$$\mathcal{D}_i^{\mathrm{src}} = \mathrm{Softmax}(W_{\mathrm{D}}(\lambda_\theta(\hat{\mathbf{v}}_i) + \mathbf{g}_i)) \tag{8}$$

$$\hat{y}_i = \mathrm{Softmax}(W_o\mathcal{D}_i^{\mathrm{src}} + b_o) \tag{9}$$

$$\mathcal{L}_{\mathrm{cross}} = -\frac{1}{\sum_{l=1}^{L}\tau(l)}\sum_{i=1}^{L}\sum_{k=1}^{\tau(i)} y_{i,k}\log(\hat{y}_{i,k}) \tag{10}$$

where L is the total number of conversations in the trainset, $\tau(i)$ is the number of utterances in the conversation, $y_{i,k}$ denotes the one-hot vector and $\hat{y}_{i,k}$ denotes probability vector for candidate emotional class n of the i^{th} utterance in l^{th} sample.

KL divergence $\mathcal{L}_{\mathrm{KL}}$ are calculated as follows:

$$\mathcal{D}_i^{\mathrm{tmp}} = \mathrm{Softmax}(W_{\mathrm{D}}\lambda_\theta(\hat{\mathbf{v}}_i)) \tag{11}$$

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL\text{-}Divergence}(\mathcal{D}_i^{tmp}, \mathcal{D}_i^{src}) \tag{12}$$

where $\lambda_\theta \in \mathbb{R}^{4d_u \times 2d_u}$ and $W_{\mathrm{D}} \in \mathbb{R}^{2d_u \times |\mathcal{E}|}$ denotes the emotional state generation network. $|\mathcal{E}|$ is the number of emotion labels. By utilizing a shared weight matrix W_{D} to map the predicted emotion $\mathcal{D}^{\mathrm{tmp}}$ and the recognized emotion $\mathcal{D}^{\mathrm{src}}$, the model is able to generate more accurate emotional representations in the current emotional state and make more precise inferences based on historical utterances.

Adversarial Emotion Disentanglement loss $\mathcal{L}_{adv}$ is proposed to further prevent the model from excessively focusing on the emotional information of a dialogue role, inspired by adversarial training methods [34]–[40]. To be more specific, given an input sentence, we obtain its hidden representations using LSTM. Next, the model classify them based on predicted probability distributions. Then, we obtain the classification cross-entropy loss L_{cross}. However, existing methods often being influenced by a specific dialogue role, it is difficult to consider the overall semantic information of the whole conversation, Therefore, we apply the Fast Gradient Value (FGV) technique [34], [35] to approximate the worst-case perturbation as a noise vector:

$$v_{noise} = \epsilon\frac{g}{\|g\|};\ \text{where } g = \nabla_e\mathcal{L}_{cross} \tag{13}$$

Here, the gradient represents the first-order derivative of the loss function $\mathcal{L}_{cross}$, and e denotes the direction of rapid increase in the loss function. We perform normalization and use a small ϵ to ensure the approximation is reasonable. Then, we add the noise vector v_{noise} and conduct a second forward pass, obtaining a new adversarial loss $\mathcal{L}'_{cross}$.

Therefore, we obtain the adversarial disentanglement loss function as follow:

$$\mathcal{L}_{adv} = \mathcal{L}_{cross} + \mathcal{L}'_{cross} \tag{14}$$

The overall training loss, namely $\mathcal{L}_{\mathrm{EC}}$ calculated as:

$$\mathcal{L}_{\mathrm{EC}} = \mathcal{L}_{\mathrm{cross}} + \alpha\mathcal{L}_{\mathrm{KL}} + \beta\mathcal{L}_{adv} \tag{15}$$

where α, β are hyperparameter mentioned in Implementation Details.

As a result, the combined loss facilitates the model's learning of emotional continuity coding and emotional attribution coding, ultimately improving its overall performance.

IV. Experiments

A. Dataset

We assess the performance of HCAN on three benchmark datasets which are IEMOCAP [26], MELD [27] and EmoryNLP [28].

IEMOCAP is a dataset recorded as dyadic conversational video clips with eight speaker participating in the training set while two speaker in testing set. Emotional tags in IEMOCAP are *happy, sad, neutral, angry, excited,* and *frustrated.*

MELD dataset is a multimodal dataset that has been expanded from the EmotionLines dataset. MELD is obtained from the popular TV show *Friends* and comprises over 1400 dialogues and 13000 utterances, each of which is labeled with emotion and sentiment classes. The emotion classes include *(i.e., happy/joy, anger, fear, disgust, sadness, surprise,* and *neutral)*, while the sentiment classes consist of *positive, negative, or neutral*.

EmoryNLP is a textual dataset also collected from the TV series *Friends*. The dataset comprises utterances that are categorized into seven distinct emotional classes, namely *neutral, joyful, peaceful, powerful, scared, mad,* and *sad*, while the sentiment classes consist of *positive, negative, or neutral*.

In this work, we only consider the emotional classes for the MELD and EmoryNLP datasets. Additionally, we maintain consistency with COSMIC in terms of the train/val/test splits. The details of datasets are presented in TABLE I and TABLE II.

TABLE I
THE STATISTICS OF SPLITS USED IN DIFFERENT DATASETS

Dataset	#Dialogue			#Utterance		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
MELD	1039	114	280	9989	1109	2610
EmoryNLP	659	89	79	7551	954	984

TABLE II
THE STATISTICS OF EVALUATION METRICS USED IN DIFFERENT DATASETS

Dataset	# classes	Metric	# Speakers
IEMOCAP	6	Weighted Avg. F1	2
EmoryNLP	7	Weighted Avg. F1	2-3
MELD	7	Weighted Avg. F1	2-3

B. Baselines

For the baselines, we mainly select two groups of outstanding models to compare with our approach.

1) Recurrent-based methods: **DialogueRNN** [12] dynamically models emotions by taking into account the current speaker, contextual content, and emotional state, with a focus on distinguishing between different speakers.

COSMIC [13] is a conversational model that incorporates commonsense knowledge to improve its performance which injects commonsense knowledge into Gated Recurrent Units to capture the internal state, external state, and intent state' features.

SKAIG [31] is improved by incorporating action information inferred from the preceding context and the intention suggested by the subsequent context. Additionally, it utilized a CSK method to represent the edges with knowledge, and introduced a graphics converter to handle them.

DialogueCRN [18] designs multi-turn reasoning modules to extract and integrate emotional clues which performs an iterative process of intuitive retrieval and conscious reasoning, mimicking the unique cognitive thinking of humans.

2) Attention-based methods: **KET** [16] utilizes external commonsense knowledge through the use of hierarchical self-attention and context-aware graph attention. This approach allows for dynamic incorporation of knowledge into transformers, resulting in a knowledge-enriched model.

DAG-ERC [32] regards the internal structure of dialogue as a directed acyclic graph to encode utterances, providing a more intuitive approach to model the information flow between the distant conversation background and the nearby context.

TODKAT [29] proposes a language model (LM) augmented with topics, which includes an additional layer specialized in detecting topics, and incorporates commonsense statements obtained from a knowledge base based on the dialogue context.

CoG-BART [20] presents a new method that employs a contrastive loss and a task for generating responses to ensure that distinct emotions are mutually exclusive.

C. Implementation Detail

Following COSMIC [13], we only utilize utterance-level text features that are fine-tuned using RoBERTa [30] to accomplish the ERC task. We conduct all HCAN experiments with a learning rate of 1e-4. The batch size is set to 32 and the dropout rate is kept at 0.2. The number of LSTM layers was set to 2, 1, and 1 on IEMOCAP, MELD, and EmoryNLP datasets, respectively. The number of heads in standard multi-head attention and IA-attention are 8 and 4, respectively. The hyperparameter α is set as 0.1, 0.2, 0.2 for IEMOCAP, MELD, and EmoryNLP datasets while β is unified as 0.05. The results reported in our experiments are based on the average score of 5 random runs on the test set. A server with one NVIDIA A100(40G) GPU is used to conduct our experiments. The addtional reproduction experiments are aligned to the baselines strictly.

D. Main Result

TABLE III shows the main results of HCAN on three benchmarks compared to previous methods. The results demonstrate that our HCAN achieves the best performance across all three

TABLE III
F1 SCORES ON THREE BENCHMARK. THE BEST RESULTS ARE IN BOLD.

Models	IEMOCAP	MELD	EmoryNLP
	W-Avg F1	W-Avg F1	W-Avg F1
KET	61.33	58.18	34.39
DialogueRNN†	62.75	-	-
TODKAT†	63.75	65.27	38.59
DialogueGCN	64.37	58.10	-
COSMIC†	65.28	64.21	37.61
DialogXL	66.2	62.41	34.73
DialogueCRN	66.33	58.39	-
SKAIG†	65.79	65.18	37.57
DAG-ERC†	68.03	63.65	38.94
COG-BART	66.18	64.81	39.04
DialogueRNN$^{\dagger}_{+\mathrm{EAE}}$	64.85(↑ 2.10)	-	-
COSMIC$^{\dagger}_{+\mathrm{EAE}}$	67.77(↑ 2.50)	65.73(↑ 1.52)	38.71(↑ 1.10)
TODKAT$^{\dagger}_{+\mathrm{EAE}}$	64.98(↑ 1.23)	65.87(↑ 0.60)	38.92(↑ 0.33)
SKAIG$^{\dagger}_{+\mathrm{EAE}}$	68.09(↑ 2.30)	65.68(↑ 0.50)	38.50(↑ 1.07)
DAG-ERC$^{\dagger}_{+\mathrm{EAE}}$	68.80(↑ 0.77)	64.73(↑ 1.08)	39.45(↑ 0.51)
HCAN(Ours)	**69.21**	**66.24**	**39.67**

† indicates our reproduction results with the same settings in baselines.

$_{+\mathrm{EAE}}$ means the model added with EAE module.

datasets. Furthermore, compared to the previous state-of-the-art (SOTA) models on IEMOCAP, MELD, and EmoryNLP, HCAN outperforms them by 1.18%, 0.95%, and 0.63%, respectively. IEMOCAP is known for having longer multi-turn dialogues and a well-balanced distribution of emotions, which allows for a more comprehensive evaluation of model performance. Our significant improvement(1.18%) in performance on this dataset successfully demonstrates the model's ability to model long-distance emotional continuity and effectiveness in dyadic conversational scenario. MELD and EmoryNLP datasets consist of multiple dialogue roles and shorter conversations, which closely resemble real-life scenarios. Additionally, these datasets have highly imbalanced emotion categories. Our model's improvement on these datasets demonstrates its effectiveness in capturing complex dialogue relationships and interpersonal emotional dependencies, as well as its robustness in recognizing different emotions. It is worth noting that the previous SOTA models were achieved using different models for each dataset, as the sample characteristics of each dataset vary significantly. However, our method unifies the SOTA across these benchmarks, demonstrating the generalizability of our approach in different application scenarios.

E. Ablation Studies

As shown in Table IV, we conducted more detailed ablation experiments to quantify the contributions of the ECE module, EAE module, $\mathcal{L}_{\mathrm{KL}}$, $\mathcal{L}_{\mathrm{sec}}$ to the performance. (1) For ECE module, the ablation experiments leads to a performance decrease of 2.75%, 0.60%and 0.32% on IEMOCAP, MELD and EmoryNLP respectively, demonstrating its generalization on different scenarios and especially effectiveness in long conversation. (2) For EAE module, the removal of EAE leads to a performance decrease of 0.67%, 1.65% and 1.99%on IEMOCAP, MELD and EmoryNLP respectively. The results elaborate the effectiveness of EAE and the importance of emotional attribution modeling based on dialogue relationship. (3) For KL loss, the removal of $\mathcal{L}_{\mathrm{KL}}$ causes a decrease in model performance by 0.93% on the EmoryNLP dataset. This suggests the effectiveness of KL in detecting emotional shifts, as this dataset often contains emotional shifting samples. Overall, the unique contributions of different modules jointly contribute to the generalization and effectiveness of HCAN.

TABLE IV
ABLATION EXPERIMENTS OF HCAN'S COMPONENTS ON THREE DIFFERENT BENCHMARKS

Models	IEMOCAP	MELD	EmoryNLP
	W-Avg F1	W-Avg F1	W-Avg F1
HCAN	**69.21**	**66.24**	**39.67**
- w/o ECE	66.46(↓ 2.75)	65.73(↓ 0.60)	38.95(↓ 0.72)
- w/o EAE	68.57(↓ 0.67)	64.59(↓ 1.65)	37.28(↓ 2.39)
- w/o $\mathcal{L}_{\mathrm{KL}}$	69.13(↓ 0.08)	65.97(↓ 0.27)	38.74(↓ 0.93)

- w/o * indicates the experimental results without the * module in HCAN

F. The Exploration of Generality

Regarding the universality of the EAE module, as it has strong transferability, we conducted experiments by adding it to different models based on recurrent and attention-based methods shown in Table II. The results show that our EAE module can effectively improve the performance of models based on different architectures. Moreover, the performance improvement on IEMOCAP, a dataset with long dialogues, is stronger than that on MELD, which has shorter conversations. Meanwhile, we observe that the improvement in models based on recurrent methods(i.e. COSMIC) is greater than that in models based on attention mechanisms(i.e TODKAT). This is logical because our EAE module is implemented based on attention mechanisms, which are naturally superior to temporal structures in modeling various levels of emotional attribution. It is reasonable to assume that attention-based methods implicitly capture emotional attribution to some extent, while our method captures more comprehensive emotional attribution information, leading to performance improvement.

G. The Robustness of Speaker Modeling

By incorporating the ECE module to capture the conversational dynamics, our model has successfully captured rich speaker characteristics. Our approach to modeling speaker robustness is primarily reflected in the **Adversarial Emotion Disentanglement** loss. To quantify the contribution of this loss

TABLE V
ABLATION EXPERIMENT OF $\mathcal{L}_{adv}$: F1 SCORES ON THREE BENCHMARK.

Models	IEMOCAP	MELD	EmoryNLP
	W-Avg F1	W-Avg F1	W-Avg F1
DialogueRNN†	62.75	-	-
TODKAT†	63.75	65.27	38.59
COSMIC†	65.28	64.21	37.61
SKAIG†	65.79	65.18	37.57
DAG-ERC†	68.03	63.65	38.94
DialogueRNN†$_{+\mathcal{L}_{adv}}$	63.61	-	-
COSMIC†$_{+\mathcal{L}_{adv}}$	66.54	65.28	38.60
TODKAT†$_{+\mathcal{L}_{adv}}$	64.12	65.54	38.78
SKAIG†$_{+\mathcal{L}_{adv}}$	67.28	65.46	38.39
DAG-ERC†$_{+\mathcal{L}_{adv}}$	68.40	64.73	39.12
HCAN(Ours)$_{-\mathcal{L}_{adv}}$	69.03	65.92	39.29
HCAN(Ours)	**69.21**	**66.24**	**39.67**

† indicates our reproduction results with the same settings in baselines.

$+\mathcal{L}_{adv}$ means the model added with $+\mathcal{L}_{adv}$.

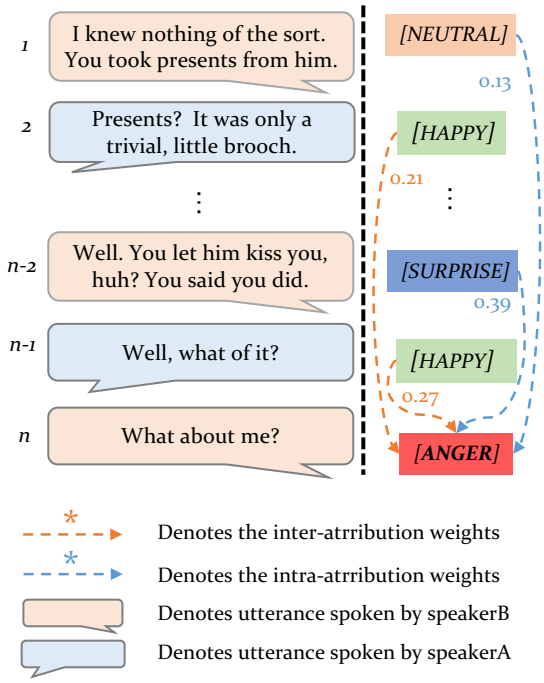

Fig. 3. Dyadic conversation for case study.

in mitigating speaker modeling overfitting, we conducted experiments similar to those in the EAE module's generalization study.

Table V shows that removing the $\mathcal{L}_{adv}$ module results in a certain degree of performance degradation for the HCAN model. Conversely, adding the $+\mathcal{L}_{adv}$ module to other baselines leads to significant performance improvements. For the SKAIG and COSMIC models, which utilize a large amount of common sense knowledge to model speaker emotions, our loss function effectively prevents overfitting on the IEMOCAP and MELD datasets, while maintaining their performance improvements. However, for models that focus on modeling conversational dynamics, such as DAG-ERC, the effect of loss improvement is limited. This is because their modeling of conversational dynamics enhances the robustness of speaker modeling to some extent.

H. Case Study

Fig. 3 shows a segment of a dyadic conversation. Intuitively, the anger expressed by speakerB in the n^{th} sentence seems to have been mainly triggered by his own surprise towards "kiss" and the question posed by speakerA in the n-1^{th} sentence. Meanwhile, the perfunctory response from speakerA in the 2^{rd} sentence may have also contributed to some extent.It is evident that the distribution of predictive emotion aligns with that of identifying the emotion, and the attention weights further demonstrate model's ability to effectively capture the relationship in long-distance conversations.

V. CONCLUSION

Insufficient modeling of context and ambiguous capture of dialogue relationships have been persistent challenges in improving the performance of ERC models. In this work, we propose HCAN to significantly addresses these issues. Our proposed ECE module achieves strong robustness in modeling global emotion continuity across different datasets by combining recurrent and attention-based approaches. It particularly demonstrates outstanding performance on long conversation samples. Meanwhile, the proposed EAE module extracts intra-attribution and inter-attribution features, which offers a more direct and accurate modeling of human emotional comprehension in the perspective of Attribution Theory. The proposed comprehensive loss function, namely Emotional Cognitive Loss $\mathcal{L}_{\mathrm{EC}}$, which effectively mitigates emotional drift and addresses the issue of overfitting in speaker modeling. Moreover, EAE module exhibits strong generalization and effectiveness when added to current models. Our model achieves state-of-the-art performance on three datasets, demonstrating the superiority of our work.

REFERENCES

[1] Munikar M, Shakya S, Shrestha A (2019) Fine-grained sentiment classifcation using bert. In: 2019 artifcial intelligence for transforming business and society (AITB), vol 1. IEEE, pp 1–5

[2] Yin D, Meng T, Chang KW (2020) Sentibert: a transferable transformer-based architecture for compositional sentiment semantics.

[3] Do HH, Prasad P, Maag A, Alsadoon A (2019) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl 118:272–299

[4] .Sun C, Huang L, Qiu X (2019) Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence.

[5] . Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," IEEE Intelligent Systems, vol. 32, no. 6, pp. 74–80, 2017.

[6] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35.

[7] Anstead. N., and O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK general election. Journal of computer-mediated communication, 20(2), 204-220.

[8] Sheridan, T. B. (2016). Human–robot interaction: status and challenges. Human factors, 58(4), 525-532.

[9] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," arXiv preprint arXiv:1811.00207, 2018.

[10] Lv, Guoqing and Wang, Xiaoping and Li, Jiang and Zeng, Zhigang. 2023. InferEM: Inferring the Speaker's Intention for Empathetic Dialogue Generation. arXiv preprint arXiv:2212.06373.

[11] Li, Jiang and Wang, Xiaoping and Lv, Guoqing and Zeng, Zhigang. 2023. GraphCFC: A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition. IEEE Transactions on Multimedia. 10.1109/TMM.2023.3260635.

[12] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 6818–6825.

[13] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020, pp. 2470–2481.

[14] . Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in Proceedings of the 2018 Conference on Empir- ical Methods in Natural Language Processing, 2018, pp. 2594–2604.

[15] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," in Proceedings of the 2019 Conference of the North American Chapter of the Association for J. Li, Z. Lin, P. Fu, and W. Wang, "Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge," in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1204–1214.

[16] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," 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), 2019, pp. 165–176.

[17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. 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 (EMNLP-IJCNLP), 2019, pp. 154–164.

[18] Hu, D., Wei, L., Huai, X. (2021, August). DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(pp. 7042-7052).

[19] W. Shen, J. Chen, X. Quan, and Z. Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 15, 2021, pp. 13 789–13 797.

[20] S. Li, H. Yan, and X. Qiu, "Contrast and generation make bart a good dialogue emotion recognizer," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11 002–11 010.

[21] Stanley Schachter and Jerome Singer. 1962. Cognitive, social and physiological determinants of emotional state. Psychological Review, 69:378–399.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, Jun. 2017.

[23] H. Chen, P. Hong, W. Han, N. Majumder, and S. Poria, "Dialogue relation extraction with document-level heterogeneous graph attention networks," arXiv preprint arXiv:2009.05092, 2020.

[24] Y. Sun, N. Yu, and G. Fu, "A discourse-aware graph neural network for emotion recognition in multi-party conversation," in Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2949–2958.

[25] Guo. M, Zhang. Y, and Liu. T. Gaussian transformer: a lightweight approach for natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 6489-6496.

[26] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N.Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, nov 2008.

[27] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. 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. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536.

[28] S. M. Zahiri and J. D. Choi, "Emotion detection on tv show transcripts with sequence-based convolutional neural networks," in The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 44–52.

[29] Zhu L, Pergola G, Gui L, et al. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 1571-1582.

[30] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in Proceedings of the 20th Chinese National Conference on Computational Linguistics. Huhhot, China: Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227.

[31] J. Li, Z. Lin, P. Fu, and W. Wang, "Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge," in Findings of the Association for Com- putational Linguistics: EMNLP 2021, 2021, pp. 1204–1214.

[32] Shen, S. Wu, Y. Yang, and X. Quan, "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 (Volume 1: Long Papers). Online: Association for Computational Linguistics, Aug. 2021, pp. 1551–1560. [Online]. Available: https://aclanthology.org/2021.acl-long.123

[33] Xue F, Sun A, Zhang H, et al. An embarrassingly simple model for dialogue relation extraction[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6707-6711.

[34] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. CoRR, abs/1412.6572

[35] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533

[36] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and A. Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In NAACL-HLT.

[37] G. Dong et al., "A Prototypical Semantic Decoupling Method via Joint Contrastive Learning for Few-Shot Named Entity Recognition," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095149.

[38] Li, X., Lei, H., Wang, L., Dong, G., Zhao, J., Liu, J., Xu, W., Zhang, C.: A robust contrastive alignment method for multi-domain text classification. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7827–7831 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747192

[39] Dong, G., Guo, D., Wang, L., Li, X., Wang, Z., Zeng, C., He, K., Zhao, J., Lei, H., Cui, X., Huang, Y., Feng, J., Xu, W.: PSSAT: A perturbed semantic structure awareness transferring method for perturbation-robust slot filling. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 5327–5334. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022), https://aclanthology.org/2022.coling-1.473

[40] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725. C. Moss-Racusin, J. F. D