Real-time Emotion Pre-Recognition in Conversations with Contrastive Multi-modal Dialogue Pre-training

Xincheng Ju
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
xcju@stu.suda.edu.cn

Dong Zhang*
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
dzhang@suda.edu.cn

Suyang Zhu
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
syzhu@suda.edu.cn

Junhui Li
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
lijunhui@suda.edu.cn

Shoushan Li
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
lishoushan@suda.edu.cn

Guodong Zhou
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, China
gdzhou@suda.edu.cn

ABSTRACT

This paper presents our pioneering effort in addressing a new and realistic scenario in multi-modal dialogue systems called **M**ulti-modal **R**eal-time **E**motion **P**re-recognition in **C**onversations (MREPC). The objective is to predict the emotion of a forthcoming target utterance that is highly likely to occur. We believe that this task can enhance the dialogue system's understanding of the interlocutor's state of mind, enabling it to prepare an appropriate response in advance. However, addressing MREPC poses the following challenges: 1) Previous studies on emotion elicitation typically focus on textual modality and perform sentiment forecasting within a fixed contextual scenario. 2) Previous studies on multi-modal emotion recognition aim to predict the emotion of existing utterances, making it difficult to extend these approaches to MREPC due to the absence of the target utterance. To tackle these challenges, we construct two benchmark multi-modal datasets for MREPC and propose a task-specific multi-modal contrastive pre-training approach[1]. This approach leverages large-scale unlabeled multi-modal dialogues to facilitate emotion pre-recognition for potential utterances of specific target speakers. Through detailed experiments and extensive analysis, we demonstrate that our proposed multi-modal contrastive pre-training architecture effectively enhances the performance of multi-modal real-time emotion pre-recognition in conversations.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; *Multimedia and multimodal retrieval*; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

multi-modal, emotion pre-recognition, contrastive learning, conversations

ACM Reference Format:
Xincheng Ju, Dong Zhang, Suyang Zhu, Junhui Li, Shoushan Li, and Guodong Zhou. 2023. Real-time Emotion Pre-Recognition in Conversations with Contrastive Multi-modal Dialogue Pre-training. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3615024

1 INTRODUCTION

Real-time emotion recognition in conversations (RERC) has become more popular than static emotion recognition (SER)[2] due to its more realistic application prospects [31, 42], such as online chatting, customer service, and dialogue systems in an ongoing environment[1, 3, 19, 27, 34]. However, in the real-time scenario of RERC, there is an interesting phenomenon: for example, in a human-machine real-time conversation, assume that the machine has responded to the user's question or request. If we wait until the user's next utterance appears, then to predict the user's emotion, we are likely to get a bad utterance (negative emotion). This is not conducive for us to evaluate the quality of the machine's response so that to modify the current reply of the machine in time if necessary. As we know, a suitable response from a machine normally provides a comfortable experience in dialogue systems [22–24].

To this end, we propose to pre-recognize the emotion of the target speaker expressed in a potential target utterance (e.g., from the user in a human-machine interactive system) yet to come, according to the historical context only. For instance, in Figure 1(a), given the distant and nearby context without target utterance, we want to pre-recognize the emotion of the target speaker for the upcoming

[*] Corresponding author
[1] https://github.com/MANLP-suda/TCMP

[2] SER typically aims at completed dialogues, considering both past and future contexts when predicting the emotion of one utterance. While, RERC can only consider the past context of the target utterance [15, 42] since the conversation is ongoing.

CIKM '23, October 21-25, 2023, Birmingham, UK
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615024

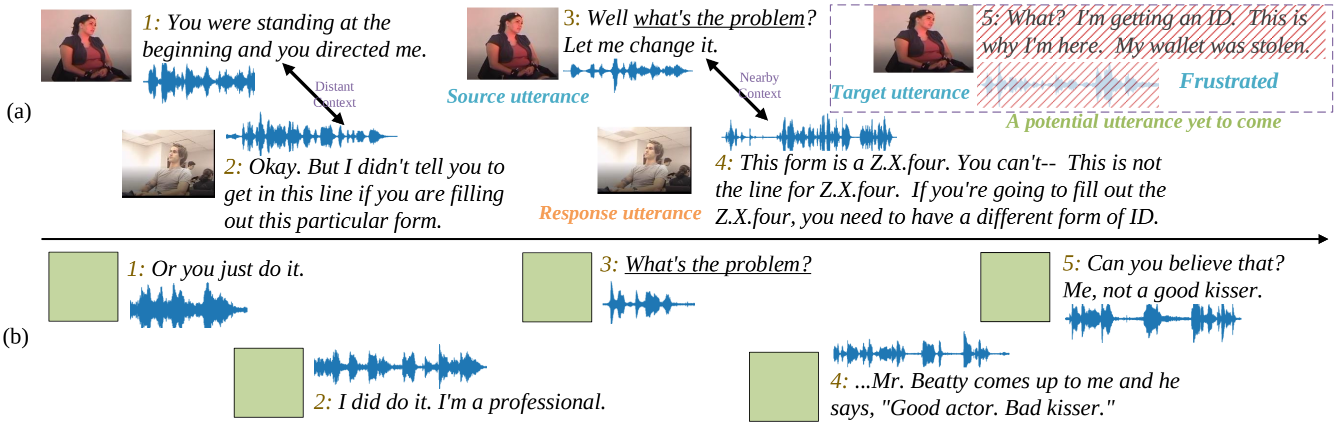

(a)
1: You were standing at the beginning and you directed me.
Distant Context
2: Okay. But I didn't tell you to get in this line if you are filling out this particular form.
3: Well what's the problem? Let me change it.
Source utterance
Nearby Context
Response utterance
4: This form is a Z.X.four. You can't-- This is not the line for Z.X.four. If you're going to fill out the Z.X.four, you need to have a different form of ID.
5: What? I'm getting an ID. This is why I'm here. My wallet was stolen.
Target utterance
Frustrated
A potential utterance yet to come
(b)
1: Or you just do it.
2: I did do it. I'm a professional.
3: What's the problem?
4: ...Mr. Beatty comes up to me and he says, "Good actor. Bad kisser."
5: Can you believe that? Me, not a good kisser.

Figure 1: (a) An example of multi-modal real-time emotion pre-recognition in conversations (MREPC). The possible target utterance 5 is yet to come, but we attempt to forecast the emotion of the target speaker expressed in the potential upcoming utterance 5 through its historical context (utterances 1-4). According to previous studies [20, 39], there is a short-term context that mostly influences the emotion of a speaker, i.e., a nearby context, which is set by us with the two utterances most close to the potential target utterance (e.g., the source utterance and response utterance, corresponding to utterances 3 and 4). To distinguish the utterances before the nearby context, we also set them as the distant context (e.g., utterances 1 and 2). (b) A dialogue segment in unlabelled multi-modal dialogues.

target utterance. In this way, the predicted emotion can help us revise the response in advance, hoping that the target speaker will have a positive emotion. Since we work in real-time scenarios with dynamic historical context and with multi-modal features in dialogue, we refer to our proposed task as **m**ulti-modal **r**eal-time **e**motion **p**re-recognition in **c**onversations (MREPC). However, for MREPC, we believe that there exist the following challenges at least:

1) Previous studies of emotion elicitation [9, 39] normally focus on textual modality only and perform sentiment forecasting in an ideal context scenario (e.g., with only two utterances as context). However, our introduced MREPC involves multiple modalities and dynamic contexts according to real-life applications. This poses a challenge to minimize the inconsistency between modalities for feature representation and adapt the dynamic context in conversations. Due to the specificity of conversational interaction, conventional task-agnostic pre-trained models for modality representation [35] are more likely to be imperfect and ineffective for our task, e.g., only using BERT [6] for conversation-level feature building. Besides, as we know, non-linguistic information (e.g., audio) has been shown as an excellent auxiliary emotional indicator [12, 43]. Therefore, we believe that a decent task-related multi-modal representation of the utterance in conversation is necessary for MREPC.

2) Previous studies of multi-modal emotion recognition [4, 13, 42] aim to predict the emotion of a given (existing) utterance. However, in our MREPC, we try to predict the emotion of the target speaker expressed in a potential upcoming utterance, which is absent in practice (e.g., the potential utterance 5 in Figure 1(a)) though we call it target utterance. This poses a great challenge to extend the traditional approaches [15, 42] to handle MREPC, since previous studies require fusing the existing target utterance and its historical context. While, in MREPC, we can only rely on the historical distant and nearby contexts with dynamic utterances. In this scenario, we can not make the target utterance attend to context and even capture the target-relevant emotional information. Thus, we believe that it is necessary to simulate a target utterance and make it related to the historic context. For example, in Figure 1(a), it’s difficult to speculate the emotion of the target utterance, due to the fact that the target utterance is not existing and the historical context can not provide adequate clues for emotion pre-recognition. While, a complete conversation though unlabelled in Figure 1(b) may greatly supply the empathic information for (a). This is mainly because the two dialogues exhibit similar semantic and emotional contexts (e.g., both own the question "what’s the problem?" and both give a negative response to the question). On this basis, if we utilize (b) to pre-train the utterance representation, we think that even if no target utterance in (a), the learned contextual representation in (a) is related to the non-existent but potential target utterance, where the knowledge of real utterance 5 in (b) may be transferred to the virtual target utterance 5 in (a).

To handle the above challenges, we propose a task-related contrastive multi-modal pre-training approach with large-scale unlabelled multi-modal dialogues, namely TCMP. This approach can not only address the task-dependent multi-modal representation but also simulate the potential target utterance to transfer emotional knowledge for better pre-recognizing the possible emotion of the target speaker. Specifically, we first design a multi-modal representation network to model historical nearby and distant context properly, then obtain the general utterance and context representation. Second, we leverage unlabelled multi-modal dialogues to make a text-audio inconsistent reduction and pre-train intra- and inter-target utterance searching by contrastive learning, which can achieve both potential target utterance simulation and task-dependent multi-modal representation. Finally, we pre-recognize the emotion of the upcoming utterance from the target speaker, based on the historical utterances with pre-trained multi-modal representation. Overall, our main contributions can be summarized as follows,

- We introduce a new multi-modal emotion analysis task, namely **m**ulti-modal **r**eal-time **e**motion **p**re-recognition in **c**onversations (MREPC). To this end, we construct two datasets for MREPC from the conventional multi-modal emotion recognition datasets IEMOCAP and MELD.
- We propose a novel task-related contrastive multi-modal pre-training (TCMP) approach with large-scale unlabelled multi-modal dialogues for MREPC. This approach can well tackle the challenges of both task-related multi-modal representation and the absence of potential target utterance.
- We conduct extensive experiments and detailed analysis on two multi-modal datasets, which show that task-dependent contrastive pre-training on unlabelled multi-modal dialogues helps build the task-related multi-modal representation and alleviates the deficiency of non-existing target utterance and the scarcity of labelled data.

2 RELATED WORK

In this section, we mainly overview the most related works about our MREPC from the following three perspectives:

2.1 Emotion Recognition in Conversations

Many studies with focus on static emotion recognition, such as [5, 8, 13, 17, 25, 29, 45, 46], leverage past and future context to speculate the emotion of the existing target utterance in the textual or multi-modal scenario. Recently, textual [7, 15] and multi-modal approaches [33, 42] observe the practical applications of real-time emotion recognition in conversations (RERC), which only leverage the past context to detect the target utterance emotion.

However, all the above studies are completely designed for emotion recognition of an existing target utterance, not emotion pre-recognition of the target speaker expressed in a possible upcoming utterance. In other words, for MREPC, we lack the target utterance and can only depend on the historical context. Therefore, we propose to only utilize historical context and try to identify the emotion of a potential upcoming utterance in a conversation.

2.2 Emotion Elicitation

Several studies [23, 24] have demonstrated that emotion elicitation can facilitate textual dialogue generation. To this end, Hasegawa et al. [9] tries to predict the emotion of the addressee and generate a response that elicits a specific emotion in the addressee's mind. Subsequently, Wang et al. [39] also set an ideal context (i.e., with only two utterances as context) to forecast the binary emotion of the third utterance.

However, these approaches not only have no ability to handle multiple modalities but also can not model the dynamic context and perform multi-class emotion prediction. Therefore, we propose to construct a multi-modal architecture and focus on forecasting the emotion of the target speaker expressed in the potential target utterance yet to come.

2.3 Multi-modal Pre-training in Dialogues

Although the pre-training strategies on text have been utilized in several studies of textual emotion analysis in dialogues [18, 37], to our best knowledge, multi-modal pre-training approaches have never been employed in multi-modal emotion analysis in dialogues. Only recently, some multi-modal pre-training approaches are introduced in multi-modal dialogue generation. Shuster et al. [36] incorporate various image fusion schemes and employ domain-adaptive pre-training and fine-tuning strategies for open-domain dialogue systems. Li et al. [21] utilize a combination of several fundamental experts to address multiple dialogue-related tasks, then introduce a pre-training schema with limited dialogue and extensive non-dialogue multi-modal data, In this way, they can enhance the understanding and generation capabilities of multi-modal dialogues.

However, all the above approaches are difficult to be directly applied in multi-modal emotion analysis regarding conversational scenarios. On the one hand, textual pre-training approaches to traditional emotion analysis can not handle the multi-modal feature fusion and the absence problem of the potential target utterance in our MREPC. On the other hand, multi-modal pre-training approaches to dialogue generation can not well achieve task-related multi-modal representation learning and the lacking problem of the possible target utterance in our MREPC as well. Therefore, we resort to external unlabeled multi-modal dialogues to enrich the multi-modal representation and ultimately simulate the potential target utterance by pre-training.

3 METHODOLOGY

In this section, we introduce our proposed TCMP approach from the following aspects: 1) Task Definition; 2) Modality Encoding and Dialogue Representation; 3) Task-related Contrastive Multi-modal Pre-training; 4) Multi-modal Real-time Emotion Pre-recognition.

3.1 Task Definition

We define the following notations, used throughout the paper. Let $\mathcal{D} = \{u_1, u_2, \cdots, u_{N-2}, u_{N-1}, u_N, [mask]_{N+1}\}$ be the set of a dialogue, where N is the number of historical utterances. Each utterance involves two utterance-aligned modalities, i.e., textual (t) and acoustic (a) modalities, which can be defined as $u_i = \{u_i^t, u_i^a\}$. u_i^t and u_i^a denote the word/frame-level sequences of text and audio, respectively.

To adapt the different influence degrees of context, we separate the historical utterances into distant C_d ($\{u_1, \cdots, u_{N-2}\}$) and nearby context C_n (e.g., source and response, corresponding to u_{N-1} and u_N), respectively. $[mask]_{N+1}$ is the potential target utterance and will not be given in the whole task, just its emotion label e_{N+1} will be set as our predicting target of MREPC. Note that the speakers of source u_{N-1} and target $[mask]_{N+1}$ utterance are the same.

The MREPC task in conversations aims to pre-recognize the emotion state y (aka. e_{N+1}) of a possible target utterance in position N+1 based on the available multi-modal historical utterances $u_{\{1:N\}}$ (aka. C_d and C_n). y belongs to a finite emotion label set $\mathcal{E}$. Our goal is to achieve the conditional probability for the emotion of the

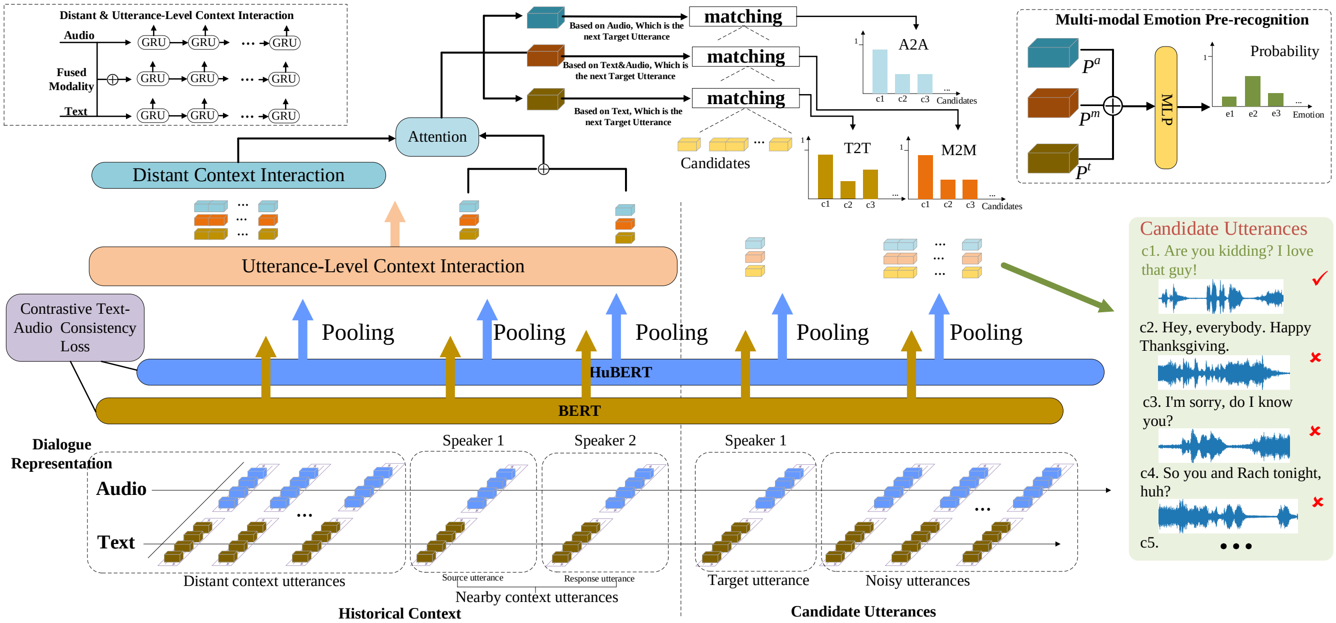

Distant & Utterance-Level Context Interaction
Audio
Fused Modality
Text
GRU
Distant Context Interaction
Attention
Based on Audio, Which is the next Target Utterance
Based on Text&Audio, Which is the next Target Utterance
Based on Text, Which is the next Target Utterance
matching
matching
matching
Candidates
A2A
T2T
M2M
1
c1
c2
c3
Candidates
Multi-modal Emotion Pre-recognition
P^a
P^m
P^t
MLP
Probability
e1
e2
e3
Emotion
Utterance-Level Context Interaction
Contrastive Text-Audio Consistency Loss
Pooling
HuBERT
BERT
Candidate Utterances
c1. Are you kidding? I love that guy!
c2. Hey, everybody. Happy Thanksgiving.
c3. I'm sorry, do I know you?
c4. So you and Rach tonight, huh?
c5.
Dialogue Representation
Audio
Text
Speaker 1
Speaker 2
Speaker 1
Distant context utterances
Source utterance
Response utterance
Nearby context utterances
Historical Context
Target utterance
Noisy utterances
Candidate Utterances

Figure 2: The overall architecture of our proposed approach TCMP. In the pre-training stage on unlabeled multi-modal dialogues, the target utterance is masked for conducting correct target utterance searching. In the fine-tuning and testing stages, the target utterance does not exist for our MREPC task.

potential $(N+1)$-th utterance:

$$p(y|\mathcal{D}) = p(e_{N+1}|C_d, C_n) \tag{1}$$

3.2 Dialogue Representation and Context Interactions

Modality Encoding. As shown in Figure 2, We first feed the text and audio utterances into two different pre-trained encoders, BERT and HuBERT. Here, BERT[6] as a language model and HuBERT[11] as an acoustic model has excellent feature extraction capability for text and audio.

Formally,

$$\begin{aligned} T_i^t &= \text{BERT}(u_i^t) \quad i \in \{1, 2, \cdots, N\} & (2)\\ T_i^a &= \text{HuBERT}(u_i^a) \quad i \in \{1, 2, \cdots, N\} & (3) \end{aligned}$$

where T_i^t and T_i^a represent the hidden state sequences of text and audio at the i-th utterance, respectively.

Utterance Representation. Then we refine the word/frame-level features with pooling and receive the utterance-level representation,

$$H_i^{\{t,a\}} = \text{MaxPool}(T_i^{\{t,a\}}) + \text{MeanPool}(T_i^{\{t,a\}}) \tag{4}$$

As above mentioned, we separate the hidden state into distant H_c and nearby (e.g., source H_s and response H_r) context representation, respectively. Formally:

$$H_c, H_s, H_r = H_{\{1:N-2\}}, H_{N-1}, H_N \tag{5}$$

Utterance-level Context Interaction. Subsequently, we fused the text and audio features and leverage GRU to capture the temporal and semantic relation among utterances as follows:

$$\begin{aligned} H^m &= W_m(H_i^t \oplus H_i^a) + b_m, \quad i \in \{1,2,\cdots,N\} & (6)\\ \hat{H}^\beta &= \text{GRU}^\beta(H^\beta) & (7) \end{aligned}$$

where $\beta = \{t, a, m\}$. $W_m \in \mathbb{R}^{2d \times d}$ and $H^m \in \mathbb{R}^{N \times d}$ is the fused modality feature.

Distant and Nearby Context Interaction. Finally, we build the context representation from the whole context for our ultimate task emotion pre-recognition. As in previous studies [20], the nearby context serves as the most direct stimulus for emotional expression. Based on this, we attempt to make the nearby context attend to the distant context, thus forming a contextual representation dominated by the nearby context.

First, we also utilize GRU to acquire emotion-driven clues in the distant context.

$$C^\beta = \text{GRU}^\beta(\hat{H}_c^\beta) \tag{8}$$

Then, we take the nearby context (concatenating the representations of the source and response utterances) as a query, and search for the relevant information in the distant context by soft attention, defined as follows:

$$\begin{aligned} Q^\beta &= (H_s^\beta \oplus H_r^\beta) W_{q\beta} + b_{q\beta} & (9)\\ P^\beta &= \text{Attention}(Q^\beta, C^\beta) & (10) \end{aligned}$$

where $W_{q\beta} \in \mathbb{R}^{2d \times d}$ is trainable parameter matrices. $P^\beta \in \mathbb{R}^{\{1,1,1\} \times d}$ are the final representation for pre-recognizing the upcoming emotion.

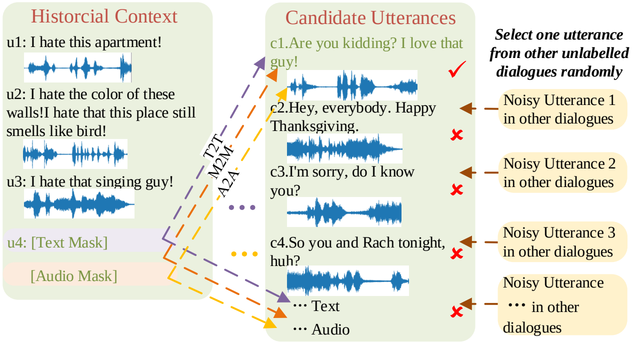
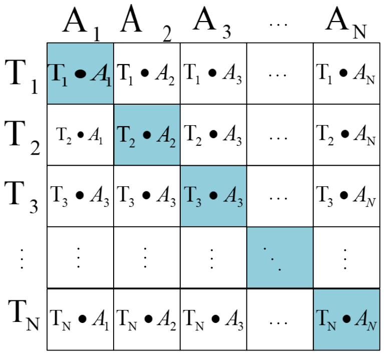

Historcial Context
u1: I hate this apartment!
u2: I hate the color of these walls!I hate that this place still smells like bird!
u3: I hate that singing guy!
u4: [Text Mask]
[Audio Mask]
T2T
M2M
A2A
Candidate Utterances
c1.Are you kidding? I love that guy!
✓
c2.Hey, everybody. Happy Thanksgiving.
✗
c3.I'm sorry, do I know you?
✗
c4.So you and Rach tonight, huh?
✗
··· Text
··· Audio
✗
Select one utterance from other unlabelled dialogues randomly
Noisy Utterance 1 in other dialogues
Noisy Utterance 2 in other dialogues
Noisy Utterance 3 in other dialogues
Noisy Utterance ··· in other dialogues

Figure 3: A sample of T2T (aka. Text2Text), A2A (aka. Audio2Audio) and M2M (aka. FusedModality2FusedModality) searching.



	A_1	A_2	A_3	$\cdots$	A_N
T_1	$T_1 \bullet A_1$	$T_1 \bullet A_2$	$T_1 \bullet A_3$	$\cdots$	$T_1 \bullet A_N$
T_2	$T_2 \bullet A_1$	$T_2 \bullet A_2$	$T_2 \bullet A_3$	$\cdots$	$T_2 \bullet A_N$
T_3	$T_3 \bullet A_3$	$T_3 \bullet A_3$	$T_3 \bullet A_3$	$\cdots$	$T_3 \bullet A_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T_N	$T_N \bullet A_1$	$T_N \bullet A_2$	$T_N \bullet A_3$	$\cdots$	$T_N \bullet A_N$

Figure 4: Contrastive Text-Audio Consistency Loss. T:Text, A:Audio

3.3 Task-related Contrastive Multi-modal Pre-training

Since the novel task MREPC lacks target utterance, we try to learn the target utterance representation by simulating in contrastive multi-modal pre-training. Specifically, we resort to the large-scale unlabelled dialogues and design a novel pre-training task by contrastive learning, namely target utterance searching (TUS). This pre-training task aims to search for the true target utterance from an utterance set (one true and multiple noise-sampled utterances). This process is inspired by negative sampling [26] and text-based pre-training study [14]. In this way, the utterance representation can absorb the external large-scale dialogue knowledge and make a directivity to the target utterance.

For details, we sample $k-1$ noise utterances elsewhere, along with the true target utterance, to form a set of k candidate utterances. The $k-1$ noise candidate utterances come from conversations in random videos, resampled at each training step. Figure 3 illustrates a sample of the searching process. As shown in the sample, given the historical utterances ($u1, u2, u3$) in conversation, the searching process is required to judge which is the true utterance among candidates at position $u4$.

Here, we separate TUS into two sub-tasks, i.e., intra- and inter-modal target utterance searching (Intra-TUS and inter-TUS) for exploiting the intra- and inter-modal interactions among utterances.

Intra-modal Target Utterance Searching. To capture the intra-modal interactions among dialogues for MREPC, we first perform intra-modal target utterance searching (Intra-TUS). Our designed sub-task Intra-TUS contains two kinds: 1) From the textual context only, searching the true target utterance among textual candidates (Text2Text); 2) From the acoustic context only, searching the true target utterance among acoustic candidates (Audio2Audio).

To obtain the candidate representation, we feed candidate utterances into encoders as formula (2-3) with pooling functions. And ultimately we get the $K^t, K^a \in \mathbb{R}^{k\times d}$, as textual and acoustic candidates representation.

To this end, we utilize the context representation $P^{\{t,a\}}$, as formula 10, to search for the most suitable one to fill the target position. To capture the matching degree, we adopt the dot-product with a sigmoid function as follows:

$$s_i^{t\to t} = \sigma((P^t)^\top K_i^t) \quad i \in \{1,2,\cdots,k\} \tag{11}$$

$$s_i^{a\to a} = \sigma((P^a)^\top K_i^a) \quad i \in \{1,2,\cdots,k\} \tag{12}$$

where $s^{t\to t}$ and $s^{a\to a}$ both $\in \mathbb{R}^{1\times k}$ denote Text2Text and Audio2Audio matching scores, respectively. $\sigma(x) = \frac{1}{1+\exp(-x)} \in (0,1)$.

Inter-modal Target Utterance Searching. To capture the inter-modal interactions among dialogues, we simultaneously perform inter-modal target utterance searching (Inter-TUS). Our designed sub-task Inter-TUS: From the fused multi-modal context only, we want to search the true fused multi-modal target among fused multi-modal candidates (FusedModality2FusedModality, M2M for short).

Similarly, the matching degree computing can be defined as follows,

$$s_i^{m\to m} = \sigma((P^m)^\top K_i^m) \quad i \in \{1,2,\ldots,k\} \tag{13}$$

where $s^{m\to m} \in \mathbb{R}^{1\times k}$. K^m is obtained by processing the candidates, as formula (2-3,6) with pooling.

Contrastive Loss of Target Utterance Searching. We adopt cross-entropy loss function for intra- and inter-TUS as follows:

$$\begin{aligned} \mathcal{L}_{Intra} &= -\log(s_i^{t\to t}\cdot s_i^{a\to a}) && (14)\\ &+ \sum_{j\in\{1,2,\cdots,k\}-i} \log(s_j^{t\to t}\cdot s_j^{a\to a}) \\ \mathcal{L}_{Inter} &= -\log(s_i^{m\to m}) && (15)\\ &+ \sum_{j\in\{1,2,\cdots,k\}-i} \log(s_j^{m\to m}) \\ \mathcal{L}_{II} &= \mathcal{L}_{Intra} + \mathcal{L}_{Inter} && (16) \end{aligned}$$

where $\{1,2,\cdots,k\}-i$ denotes the first set excluding the element i. Here, i is the position of the true target utterance in candidates.

Contrastive Loss of Text-Audio Consistency. Since the text and audio exhibit inconsistent semantic spaces, we try to alleviate the gap between them. Inspired by CLIP [32], we further define the contrastive loss to bridge the consistency between text and audio. Figure 4 illustrates the building status of this loss, which is

Table 1: The statistics of our reconstructed MREPC datasets, i.e., P-IEM and P-MELD. Fru: frustrated, Exc: excited, Neu: neutral, Ang: angry, Sad: sad, Hap: happy, oth: other Pos: positive, Neg: negative

Dataset	P-IEM										P-MELD					
	NU	ADL	AWL	Fru	Exc	Neu	Ang	Sad	Hap	oth	NU	ADL	AWL	Pos	Neg	Neu
train	4512	19.5	13.1	731	864	348	521	1335	592	121	4687	4.7	9.9	1009	1536	2142
dev	1737	19.8	13.7	401	215	57	109	592	317	46	506	4.7	10.0	111	199	196
test	1861	19.8	13.9	375	341	131	241	549	180	44	1178	4.7	9.9	237	405	536

Table 2: The statistics of the unlabelled multi-modal dialogues dataset (UMD). NU: The number of utterances, ADL: Avg. dialogue length, AWL: Avg. word length

Dataset	UMD		
	NU	ADL	AWL
Train	74,427	363.1	8.8
Dev	5,438	362.5	8.8
Test	5,039	359.0	8.9

formulated as follows:

$$logits = H^t \cdot (H^a)^\top \tag{17}$$

$$\mathcal{L}_h = -\sum_{i=1}^{N} \log \frac{\exp^{logits[i,i]}}{\sum_{j=1}^{N} \exp^{logits[i,j]}} \tag{18}$$

$$\mathcal{L}_v = -\sum_{i=1}^{N} \log \frac{\exp^{logits[i,i]}}{\sum_{j=1}^{N} \exp^{logits[j,i]}} \tag{19}$$

$$\mathcal{L}_{hv} = (\mathcal{L}_h + \mathcal{L}_v)/2 \tag{20}$$

where $logits \in \mathbb{R}^{N \times N}$ is the consistent scores. $\mathcal{L}_h$ is the horizontal loss and $\mathcal{L}_v$ is the vertical loss.

Total Pre-training Loss. We combine the above two losses together as follows:

$$\mathcal{L}_{total} = \zeta \mathcal{L}_{II} + \eta \mathcal{L}_{hv} \tag{21}$$

where ζ and η are two weighted hyper-parameters.

3.4 Multi-modal Real-time Emotion Pre-recognition

After contrastive multi-modal pre-training on unlabelled multi-modal dialogues, we start transferring the pre-trained task-related representation to our multi-modal real-time emotion pre-recognition task (MREPC).

Pre-recognition. As shown in Figure 2, we append a multi-layer perceptron network (aka MLP) followed by a $softmax$ function to pre-recognize the probability distribution of emotions, defined as follows:

$$\widetilde{U} = \rho(\varrho((P^t \oplus P^m \oplus P^a)W_1 + b_1)W_2 + b_2)) \tag{22}$$

where $W_1 \in \mathbb{R}^{3d \times d}$ and $W_2 \in \mathbb{R}^{d \times |\mathcal{E}|}$. ρ and ϱ denote the activation function $softmax$ and $Tanh$ respectively. $\widetilde{U} \in \mathbb{R}^{1 \times |\mathcal{E}|}$ denotes the probability of each category and $|\mathcal{E}|$ is the number of emotion categories.

Loss of MREPC. Since the distribution of categories is imbalanced, we adopt a weighted categorical cross-entropy loss function to optimize the model parameters:

$$\mathcal{J} = -\omega(y) \cdot log(\widetilde{U}^y) \tag{23}$$

where $\widetilde{U}^y$ is the predicted probability of true emotion y. The weight vector ω is inversely proportional to the ratio of emotion categories and $\omega(y)$ is the weight of emotion y.

4 EXPERIMENTATION

In this section, we conduct systematical experiments to evaluate our proposed TCMP approach to MREPC task on two newly-constructed datasets.

4.1 Datasets

Dataset. There exist two kinds of data we used in this paper: unlabelled data for pre-training and labelled data for MREPC.

Unlabelled Multi-modal Dialogues. Our unlabelled multi-modal dialogues data is collected from TV series Friends, namely UMD. Each dialogue contains many utterances, and each utterance contains text and audio by separating the video. Table 2 shows the statistics of the unlabelled multi-modal dialogues data. We pre-train the model with 74427 utterances, which is far beyond the length of the labelled dialogue dataset.

Muti-modal Emotion Pre-recognition Datasets. We evaluate our proposed approach TCMP on two multi-modal dialogue emotion datasets, namely P-IEM and P-MELD. 1) P-IEM comes from IEMOCAP[2]: A multimodal conversational dataset for emotion recognition, which includes nine classes: neutral, happy, sad, angry, frustrated, surprised, fear, disgusted and excited. 2) P-MELD comes from MELD [30]: A multimodal dataset for emotion recognition collected from the TV show Friends, which contains three coarse-grained classes: positive, neutral, negative. To adapt our emotion pre-recognition task, we reconstruct both datasets. We control the speaker of the source utterance speaker to be the same as that of the target utterance and set the emotion of target utterance from target speaker as the final label for pre-recognition. The complete example can refer to Figure 1(a). Note that the target utterance in each sample is masked. The statistics summary of them are shown in Table 1.

4.2 Experimental Settings

Implementation Details. We implement our approach via Pytorch toolkit (torch-1.1.0) with a piece of GeForce RTX 3090. The hidden size d in our model is 768. Besides, we utilize the dropout regularization [38] to avoid overfitting and set the maximum norm of gradient clipping [28] as 5.0. During training, we set the epochs of model as 50 and test its performance on the validation set. Once the training is finished, we choose the model with the best WA score on the validation set and evaluate its performance on the test set. We use the Adam [16] optimization method to minimize the loss over the training data. For the hyper-parameter of Adam optimizer, we have the following settings: 1) For contrastive pre-training, we set the learning rate as 2e-6 for pre-trained Modules BERT and HuBERT and 2e-4 for other parameters. 2) for fine-tuning, we set the learning rate as 2e-6 for pre-trained Modules BERT and HuBERT and 2e-5 for other parameters. The two step training Adam optimizer contains two momentum parameters of β_1 and β_2, 0.9 and 0.999 respectively. For the hyper-parameter setting of loss, $\zeta = 0.5$ and $\eta = 0.5$.

Evaluation Metrics. To evaluate the performance of pre-training, we report the top-k ranking performance on TUS, following the previous pre-training approaches [14, 40, 47]. And we adopt the mainstream evaluation metrics: R5@1, R5@2, R11@1, and R11@2. Formally,

$$R_k@m = \frac{\sum_{i=1}^{m} \xi_i}{\sum_{i=1}^{k} \xi_i} \tag{24}$$

where $R_k@m$ is the recall of the true positives among m best-matched answers from k available candidates for the potential target utterance in a conversation. The variable ξ_i represents the binary label for each candidate, i.e., 1 for the correct one and 0 for the noisy ones. .

To evaluate our main task MREPC, we refer to previous works of ERC [8, 10, 25] and emotion elicitation [9, 39] by adopting the macro-averaged F_1 score, average accuracy (AA) and weighted accuracy (WA).

4.3 Baselines

For a thorough comparison, we also implement the state-of-the-art (SOTA) of multi-modal ERC by adjusting them to our pre-recognition task, besides the approach of emotion elicitation. In our implementation of these SOTA, we cast the nearby context as the target utterance and the distant context as the complete historical context. We believe that the primary motivation for this modification is consistent with our approach.

- **PEE** [9] mainly aims to investigate two novel tasks: predicting the emotion of the addressee and generating a response that elicits a specific emotion in the addressee's mind. For emotion prediction, the paper constructs a one-versus-the-rest classifier by combining eight binary classifiers. Considering this approach conducted only on textual modality, we fuse the multiple modalities as input. Besides, we also report the results on text only, with the marker **PEE-Text**.
- **NSF** [39] propose a Neural Sentiment Forecasting (NSF) model to address inherent challenges of target utterance unknown (yet to come). Note that NSF only serves in text-based conversation scenarios with an ideal context. In our implementation, we append a powerful encoder (i.e., BERT and HuBERT) to the model and fuse the multiple modalities as input. Then we mark the results with **NSF-our impl.** Besides, we also report the results on text only, with the marker **NSF-Text-our impl.**.

Table 3: The performance of target utterance searching (TUS) in pre-training with different perspectives. The intra-modal part comes from our model which only calculates the intra-modal TUS loss (Eq. 14), while the inter-modal part only calculates the inter-modal TUS loss (Eq. 15). Multi-modal part denotes that both the intra- and inter-modal TUS loss are calculated simultaneously (Eq. 16). Note that higher scores mean better experimental results. T2T: Text2Text, A2A: Audio2Audio, M2M: FusedModality2FusedModality.

Approaches	**TUS**	$R_5@1$	$R_5@2$	$R_{11}@1$	$R_{11}@2$
Intra-modal	T2T	41.18	63.74	24.60	40.13
	A2A	96.38	97.20	96.89	96.93
Inter-modal	M2M	80.82	82.79	79.96	80.91
Multi-modal	T2T	45.36	67.12	27.88	41.70
	A2A	87.57	89.59	86.05	87.50
	M2M	90.59	92.84	89.33	90.46

- **AGHMN** [15] proposes a gated hierarchical memory network for real-time emotion recognition in conversations. Note AGHMN only serves in text-based conversation scenarios. In our implementation, we fuse the multiple modalities as input. Besides, we also report the results on text only, with the marker **AGHMN-Text**.
- **DialogXL** [35] proposes an all-in-one XLNet model for multi-party conversation emotion recognition, with enhanced memory to store longer historical context and dialog-aware self-attention to deal with the multi-party structures. DialogXL still serves in text-based conversation and is applied in real-time emotion recognition. We fuse the multiple modalities as input and report the results with the marker **DialogXL**. Meantime, we also report the results on text only, with the marker **DialogXL-Text**. In our implementation, we adjust the parameter settings and make the pre-trained module XLNet adapted to the new task. Then we mark the results with **DialogXL-our impl.**
- **BiDDIN** [42] proposes a bidirectional dynamic dual influence network for real-time emotion detection in conversations, which can simultaneously model both intra- and inter-modal influence.
- **MMGCN** [13] proposes a new model based on a multi-modal fused graph convolutional network for utilizing both multi-modal and long-distance contextual information effectively.
- **MDI** [44] propose a Multi-modal Multi-scene Multi-label Emotional Dialogue dataset and a general Multimodal Dialogue-aware Interaction framework to model the dialogue context for emotion recognition. In our implementation, we append a powerful encoder (i.e., BERT and HuBERT) to the model and adjust the parameter settings. Then, we mark the results with **MDI-our impl.**

4.4 Experimental Results

Pre-training Results. Table 3 shows the performance of our pre-training. on the test set of unlabelled multi-modal dialogues data(UMD). The results show that our multi-modal pre-training approach outperforms intra-modal approaches in some metrics, especially in T2T and M2M. For instance, $R_5@1$ and $R_5@2$ of T2T in multi-modal approach increase significantly by 4.2% and 3.4% compared to that of intra-modal. While, audio related scores perform

Table 4: The performance of different approaches for MREPC task. Text: only utilize textual modality. our impl.: implementing our modified setting for the corresponding approach.

Modality	Approaches	P-MELD			P-IEM		
		WA	**AA**	**F1**	**WA**	**AA**	**F1**
Text	PPE-Text	45.5	33.3	20.9	29.4	10.0	4.5
	NSF-Text-our impl.	48.6	41.3	35.6	45.2	17.9	12.0
	AGHMN-Text	43.4	36.4	32.1	41.7	20.7	19.4
	DialogXL-Text	45.5	33.3	20.9	44.9	20.5	18.8
Text+Audio	PPE[9]	43.3	33.9	29.2	34.8	12.6	9.3
	NSF-our impl. [39]	49.2	41.3	36.4	44.4	17.5	11.7
	AGHMN [15]	44.7	33.7	30.5	35.4	13.1	9.5
	DialogXL [35]	42.4	35.3	31.3	45.3	21.1	18.3
	DialogXL-our impl.	46.3	36.9	32.5	48.8	27.9	26.4
	BiDDIN [42]	42.2	34.4	30.5	34.3	14.7	12.7
	MMGCN [13]	46.5	34.6	24.7	36.9	22.8	20.5
	MDI-our impl. [44]	46.7	40.0	35.2	42.6	17.3	11.8
	TCMP(ours)	**50.1**	**41.5**	**38.9**	**53.9**	**28.2**	**27.4**

Table 5: The performance of single-modal and multi-modal ablated approaches on both datasets. -Pre denotes the pre-trained approach. w/o Dist denotes the elimination of distant context.

Approaches	P-MELD			P-IEM		
	WA	**AA**	**F1**	**WA**	**AA**	**F1**
Text w/o Dist	49.2	41.2	36.7	44.8	17.8	11.9
Text	49.5	41.7	36.9	44.6	17.6	11.8
Text-Pre	49.3	41.1	37.8	48.9	21.2	18.8
Text+Audio w/o Dist	48.4	41.5	37.1	46.0	19.6	15.8
Text+Audio	49.6	42.1	36.8	52.7	27.3	26.8
TCMP(ours)	**50.1**	41.5	**38.9**	**53.9**	**28.2**	**27.4**

better than text-related counterparts, e.g., A2A better than M2M, and greatly better than T2T. This is mainly because audio contains some phonological clues, such as the same timbre of the speaker and the tone of voice, supporting the judgment of which is the next (target) utterance. These improvements in the multi-modal approach demonstrate that training with both intra- and inter-modal interaction simultaneously can help modalities complement each other and our multi-modal pre-training strategy is meaningful and applicable.

Emotion Pre-recognition Results. Table 4 shows the results of different approaches for MREPC on uni-modal and multi-modal data.

For uni-modal(aka textual) approaches, we observed that some text-based baselines perform worse than their multi-modal counterparts, e.g., **PPE**, but some perform the opposite, e.g., **AGHMN**. This is mainly because acoustic modality may not only enrich the semantic information, but sometimes also brings much noise to textual modality inevitably.

For multi-modal approaches, we observe that 1) multi-modal approach **MMGCN** performs better than the modality concatenating approaches **AGHMN** and **PPE** (originally text-domain approaches) in most metrics. This is mainly because the multi-modal approach **MMGCN** has advantages to dealing with multiple modalities, compared to uni-modal(e.g, textual) approaches. 2) Though **DialogXL** is not applicable to multi-modal scenario, the pre-trained module XLNet[41] in **DialogXL** still have its efficiency. We modified the approach and implement our parameters, finding that **DialogXL-our impl.** performs much better than **MLGCN**. 3) Multi-modal approach **BiDDIN** performs worse than **AGHMN** and **DialogXL**. This clearly reveals that **BiDDIN** may not have the ability to handle non-existing target utterance in MREPC task. 4) Among all the approaches, our proposed **TCMP** performs the best significantly. Especially on P-IEM, **TCMP** outperforms **DialogXL-our impl.** by 5.1%, 0.3%, 1.0% with respect to WA, AA and F_1 respectively. We speculate that there are some reasons as follows: 1) Our approach minimizes the inconsistency between modalities and adapts dynamic context in conversations. 2) We absorb some external knowledge from large-scale unlabelled multi-modal dialogues and learn to simulate a target utterance related to the potential upcoming utterance by contrastive pre-training.

5 ANALYSIS

In this section, we give a further investigation of some experimental results in Table 5.

Ablation Study. 1) To illustrate the significance of multiple modalities, we compare the multi-modal approaches and their uni-modal counterparts upon two datasets. Here, we clearly find that multi-modal approaches perform better than uni-modal approaches. This mainly because our approach can help modalities complement each other and have a strong ability of multi-modal representation constructing.

2) To illustrate the impact of our contrastive pre-training method deeply, we compare the pre-trained approaches and their non-pre-trained counterparts. From this table , we observe that pre-trained approaches perform better than their non-pre-trained counterparts significantly in the textual or multi-modal scenario. For instance on dataset P-IEM, Text-Pre outperforms Text by 4.3%, 3.6%, and

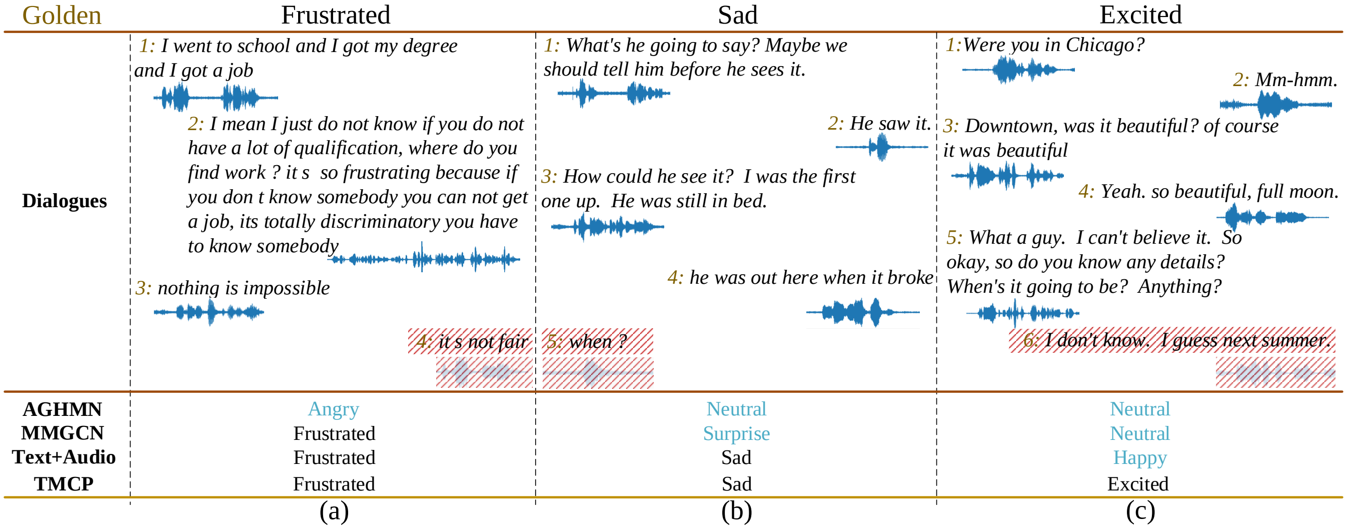

Golden
Frustrated
Sad
Excited
Dialogues
1: I went to school and I got my degree and I got a job
2: I mean I just do not know if you do not have a lot of qualification, where do you find work ? it s so frustrating because if you don t know somebody you can not get a job, its totally discriminatory you have to know somebody
3: nothing is impossible
4: it s not fair
1: What's he going to say? Maybe we should tell him before he sees it.
2: He saw it.
3: How could he see it? I was the first one up. He was still in bed.
4: he was out here when it broke
5: when ?
1:Were you in Chicago?
2: Mm-hmm.
3: Downtown, was it beautiful? of course it was beautiful
4: Yeah. so beautiful, full moon.
5: What a guy. I can't believe it. So okay, so do you know any details? When's it going to be? Anything?
6: I don't know. I guess next summer.
AGHMN
MMGCN
Text+Audio
TMCP
Angry
Frustrated
Frustrated
Frustrated
Neutral
Surprise
Sad
Sad
Neutral
Neutral
Happy
Excited
(a)
(b)
(c)

Figure 5: The predictions of three cases by AGHMN, MMGCN, Text+Audio, and TCMP.

7.0% with respect to WA, AA and F_1 respectively. While, TCMP increases 1.2%, 0.9%, and 0.6% compared with Text+Audio. This illustrates that pre-training with large-scale unlabelled dialogues enriches the modality representation and promotes semantic and empathic expression.

3) To illustrate the necessity of distant context, we ablate the distant context. From this table, we observe that removing the distant context significantly decreases the performance. Especially on P-IEM, Text+Audio w/o Dist performs lower than Text+Audio by 6.7%, 7.7% and 11.0% intuitively. This indicates that the distant context incorporates much empathic information. Our approach is capable of only utilizing historical utterances (e.g, distant and nearby context) in a dynamic context and can excellently capture the empathic relevant information in distant context with the assistance of nearby context.

Case Study. We present three cases in Figure 5. We can obviously realize that: In example (a), AGHMN gives a wrong emotion prediction of "Angry", owing to the disability of handling multi-modal dynamics, i.e., intra- and inter-modal interactions. In example (b), AGHMN and MMGCN both predict an error aspect. This is mainly because both approaches are devoted to the emotion recognition task, not to the MREPC task, which can not handle the dynamic context well without the assistance of the target utterance. In example (c), we found that all the approaches except TMCP predict incorrectly, suggesting that without pre-training to simulate the target utterance and absorb some external semantic information from large-scale unlabelled dialogues, the performance decreases obviously.

6 CONCLUSION

In this paper, we propose a task-related multi-modal contrastive pre-training approach with large-scale unlabelled multi-modal dialogues to emotion pre-recognition for the potential utterance of a specific target speaker. Our approach can not only model the multiple modalities and dynamic context, but also wisely make a task-related pre-training to simulate the non-existing target utterance. Besides, **TCMP** absorbs external unlabelled dialogue knowledge, alleviating the scarcity of labelled pre-recognition datasets.

In our further work, we will enlarge our unlabelled multi-modal dialogue data for better pre-training performance and extend our approach to more multi-modal dialogue scenarios, such as intentional understanding and conversation management. Furthermore, we would like to investigate other approaches (e.g., reinforcement learning and few-shot learning) to better improve the performance of MREPC.

7 ACKNOWLEDGMENTS

This work is supported by NSFC grants No.62206193, No. 62076176, and No.62076175.

REFERENCES

[1] Al-Hanouf Al-Aljmi and Nora Al-Twairesh. 2021. Building an Arabic Flight Booking Dialogue System Using a Hybrid Rule-Based and Data Driven Approach. *IEEE Access* 9 (2021), 7043–7053. https://doi.org/10.1109/ACCESS.2021.3049732

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation* 42, 4 (2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[3] Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, and Mihir Sanjay Kale. 2021. TicketTalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *Proceedings of ACL 2021*. Association for Computational Linguistics, 671–680. https://doi.org/10.18653/v1/2021.acl-long.55

[4] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao. 2021. Learning What and When to Drop: Adaptive Multimodal and Contextual Dynamics for Emotion Recognition in Conversation. In *Proceedings of ACM MM 2021*. 1064–1073. https://doi.org/10.1145/3474085.3475661

[5] Iek-Heng Chu, Ziyi Chen, Xinlu Yu, Mei Han, Jing Xiao, and Peng Chang. 2022. Self-supervised Cross-modal Pretraining for Speech Emotion Recognition and Sentiment Analysis. In *Findings of EMNLP 2022*. Association for Computational Linguistics, 5105–5114. https://aclanthology.org/2022.findings-emnlp.375

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[7] Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Proceedings of EMNLP 2020 (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 2470–2481. https://doi.org/10.18653/v1/2020.findings-emnlp.224

[8] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of EMNLP-IJCNLP 2019*. Association for Computational Linguistics, 154–164. https://doi.org/10.18653/v1/D19-1015

[9] Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and Eliciting Addressee's Emotion in Online Dialogue. In *Proceedings of ACL 2013*. The Association for Computer Linguistics, 964–972. https://aclanthology.org/P13-1095/

[10] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of EMNLP 2018*. Association for Computational Linguistics, 2594–2604. https://doi.org/10.18653/v1/d18-1280

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

[12] Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. *CoRR* abs/2203.02385 (2022). https://doi.org/10.48550/arXiv.2203.02385 arXiv:2203.02385

[13] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of ACL/IJCNLP 2021*. Association for Computational Linguistics, 5666–5675. https://doi.org/10.18653/v1/2021.acl-long.440

[14] Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Exploiting Unsupervised Data for Emotion Recognition in Conversations. In *Proceedings of EMNLP 2020 (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 4839–4846. https://doi.org/10.18653/v1/2020.findings-emnlp.435

[15] Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In *Proceedings of AAAI 2020*. AAAI Press, 8002–8009. https://aaai.org/ojs/index.php/AAAI/article/view/6309

[16] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR 2015*. http://arxiv.org/abs/1412.6980

[17] Joosung Lee and Wooin Lee. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In *Proceedings of NAACL 2022*. Association for Computational Linguistics, 5669–5679. https://doi.org/10.18653/v1/2022.naacl-main.416

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of ACL 2020*. Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[19] Lin Li, Wanzhong Zhao, Can Xu, Chunyan Wang, Qingyun Chen, and Shijuan Dai. 2021. Lane-Change Intention Inference Based on RNN for Autonomous Driving on Highways. *IEEE Trans. Veh. Technol.* 70, 6 (2021), 5499–5510. https://doi.org/10.1109/TVT.2021.3079263

[20] Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng. 2018. Inferring User Emotive State Changes in Realistic Human-Computer Conversational Dialogs. In *Proceedings of ACM MM 2018*. 136–144. https://doi.org/10.1145/3240508.3240575

[21] Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. 2023. PaCE: Unified Multi-modal Dialogue Pre-training with Progressive and Compositional Experts. In *Proceedings of ACL 2023*.

[22] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of LREC 2016*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/947.html

[23] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach. In *Proceedings of AAAI 2018*. AAAI Press, 5293–5300. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16317

[24] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Positive Emotion Elicitation in Chat-Based Dialogue Systems. *IEEE ACM Trans. Audio Speech Lang. Process.* 27, 4 (2019), 866–877. https://doi.org/10.1109/TASLP.2019.2900910

[25] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceeding of AAAI 2019*. AAAI Press, 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

[26] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS 2013*. 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[27] Desmond C. Ong, Marie Therese Quieta, and Basura Fernando. 2020. Intention Inference in a Dynamic Multi-Goal Environment. In *Proceedings of CogSci 2020*. cognitivesciencesociety.org. https://cogsci.mindmodeling.org/2020/papers/0392/index.html

[28] Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of ICML 2013 (JMLR Workshop and Conference Proceedings, Vol. 28)*. JMLR.org, 1310–1318. http://proceedings.mlr.press/v28/pascanu13.html

[29] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of ACL 2017*. Association for Computational Linguistics, 873–883. https://doi.org/10.18653/v1/P17-1081

[30] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of ACL 2019*. Association for Computational Linguistics, 527–536. https://doi.org/10.18653/v1/p19-1050

[31] Libo Qin, Wanxiang Che, Minheng Ni, Yangming Li, and Ting Liu. 2021. Knowing Where to Leverage: Context-Aware Graph Convolutional Network With an Adaptive Fusion Layer for Contextual Spoken Language Understanding. *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), 1280–1289. https://doi.org/10.1109/TASLP.2021.3053400

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML 2021 (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[33] Sandratra Rasendrasoa, Alexandre Pauchet, Julien Saunier, and Sébastien Adam. 2022. Real-Time Multimodal Emotion Recognition in Conversation for Multi-Party Interactions. In *Proceedings of ICMI 2022*. ACM, 395–403. https://doi.org/10.1145/3536221.3556601

[34] Michal Satlawa, Katarzyna Zamlynska, Jaroslaw Piersa, Joanna Kolis, Klaudia Firlag, Katarzyna Beksa, Zuzanna Bordzicka, Christian Goltz, Pawel Bujnowski, and Piotr Andruszkiewicz. 2021. SRPOL DIALOGUE SYSTEMS at SemEval-2021 Task 5: Automatic Generation of Training Data for Toxic Spans Detection. In *Proceedings of ACL 2021*. Association for Computational Linguistics, 974–983. https://doi.org/10.18653/v1/2021.semeval-1.133

[35] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *Proceedings of AAAI 2021*. AAAI Press, 13789–13797. https://ojs.aaai.org/index.php/AAAI/article/view/17625

[36] Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-Modal Open-Domain Dialogue. In *Proceedings of EMNLP 2021*. Association for Computational Linguistics, 4863–4883. https://doi.org/10.18653/v1/2021.emnlp-main.398

[37] Tiberiu Sosea and Cornelia Caragea. 2021. eMLM: A New Pre-training Objective for Emotion Related Tasks. In *Proceedings of ACL/IJCNLP 2021*. 286–293. https://doi.org/10.18653/v1/2021.acl-short.38

[38] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958. http://dl.acm.org/citation.cfm?id=2670313

[39] Zhongqing Wang, Xiujun Zhu, Yue Zhang, Shoushan Li, and Guodong Zhou. 2020. Sentiment Forecasting in Dialog. In *Proceedings of COLING 2020*. 2448–2458. https://doi.org/10.18653/v1/2020.coling-main.221

[40] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of ACL 2017*. Association for Computational Linguistics, 496–505. https://doi.org/10.18653/v1/P17-1046

[41] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of NeurIPS 2019,*. 5754–5764. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

[42] Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Modeling both Intra- and Inter-modal Influence for Real-Time Emotion Detection in Conversations. In *Proceedings of ACM MM 2020*. 503–511. https://doi.org/10.1145/3394171.3413949

[43] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. 2021. MEmoBERT: Pre-training Model with Prompt-based Learning for Multimodal Emotion Recognition. *CoRR* abs/2111.00865 (2021). arXiv:2111.00865 https://arxiv.org/abs/2111.00865

[44] Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In *Proceedings of ACL 2022*. Association for Computational

Linguistics, 5699–5710. https://doi.org/10.18653/v1/2022.acl-long.391

[45] Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In *Proceedings of IJCAI 2022*. ijcai.org, 4524–4530. https://doi.org/10.24963/ijcai.2022/628

[46] Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022. MuCDN: Mutual Conversational Detachment Network for Emotion Recognition in Multi-Party Conversations. In *Proceedings of COLING 2022*. International Committee on Computational Linguistics, 7020–7030. https://aclanthology.org/2022.coling-1.612

[47] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *Proceedings of ACL 2018*. Association for Computational Linguistics, 1118–1127. https://doi.org/10.18653/v1/P18-1103