

A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations

Wenjie Zheng¹, Jianfei Yu^{1*}, Rui Xia^{1*}, and Shijin Wang^{2,3}

¹School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China

²FLYTEK AI Research (Central China)

³State Key Laboratory of Cognitive Intelligence, Hefei, China

¹{wjzheng, jfyu, rxia}@njust.edu.cn, ^{2,3}sjwang3@iflytek.com

Abstract

Multimodal Emotion Recognition in Multi-party Conversations (MERMC) has recently attracted considerable attention. Due to the complexity of visual scenes in multi-party conversations, most previous MERMC studies mainly focus on text and audio modalities while ignoring visual information. Recently, several works proposed to extract face sequences as visual features and have shown the importance of visual information in MERMC. However, given an utterance, the face sequence extracted by previous methods may contain multiple people’s faces, which will inevitably introduce noise to the emotion prediction of the real speaker. To tackle this issue, we propose a two-stage framework named **Facial** expression-aware **Multimodal Multi-Task** learning (FacialMMT). Specifically, a pipeline method is first designed to extract the face sequence of the real speaker of each utterance, which consists of multimodal face recognition, unsupervised face clustering, and face matching. With the extracted face sequences, we propose a multimodal facial expression-aware emotion recognition model, which leverages the frame-level facial emotion distributions to help improve utterance-level emotion recognition based on multi-task learning. Experiments demonstrate the effectiveness of the proposed FacialMMT framework on the benchmark MELD dataset. The source code is publicly released at <https://github.com/NUSTM/FacialMMT>.

1 Introduction

Multimodal Emotion Recognition in Multi-party Conversations (MERMC) is a challenging task in the field of multimodal research. The complexity of the task arises from the dynamic and spontaneous nature of human communication in multi-party conversations, which often involves multiple people expressing a variety of emotions simultaneously. In this task, the use of multiple modalities (e.g.,

Methods	#Col	#Seg	#Rec
MELD (Poria et al., 2019)	✗	✗	✗
UniMSE (Hu et al., 2022b)	✗	✗	✗
MMGCN (Hu et al., 2021)	✓	✗	✗
MESM (Dai et al., 2021)	✓	✗	✗
M ³ ED (Zhao et al., 2022b)	✓	✗	✗
FacialMMT (Ours)	✓	✓	✓

Table 1: Comparison between different models for face sequence extraction in the MERMC task. #Col represents collection of all possible speakers’ face sequences; #Seg represents speaker segmentation, aiming to distinguish speaker sequences; #Rec represents speaker recognition, aiming to identify the real speaker.

text, audio, and vision) is essential as it allows for a more comprehensive understanding of the emotions being expressed. Among different modalities, visual information usually plays a crucial role as it often provides direct clues for emotion prediction. For example, in Figure 1, without the information from the visual modality, it is hard to determine the *anger* emotion of the real speaker, i.e., *Chandler*.

In the literature, most previous MERMC studies primarily focus on the text and audio modalities (Poria et al., 2019; Liang et al., 2020; Mao et al., 2021; Chen et al., 2021), because the visual context in MERMC often involves many people and complex environmental scenes, which may bring much noise to emotion recognition of the real speaker. Owing to the indispensable role of visual modalities, a number of studies explored the potential of visual information in MERMC (Mai et al., 2019; Wang et al., 2022a; Li et al., 2022b; Hu et al., 2022a), which employ 3D-CNNs (Ji et al., 2010; Tran et al., 2015) to extract video features and model the interaction and dependency between consecutive video frames. However, the visual information extracted by these methods still contains much noise from environmental scenes.

To alleviate the visual noise from environmental scenes, several recent studies (Dai et al., 2021; Liang et al., 2021; Hu et al., 2021; Zhao et al.,

* Corresponding authors.

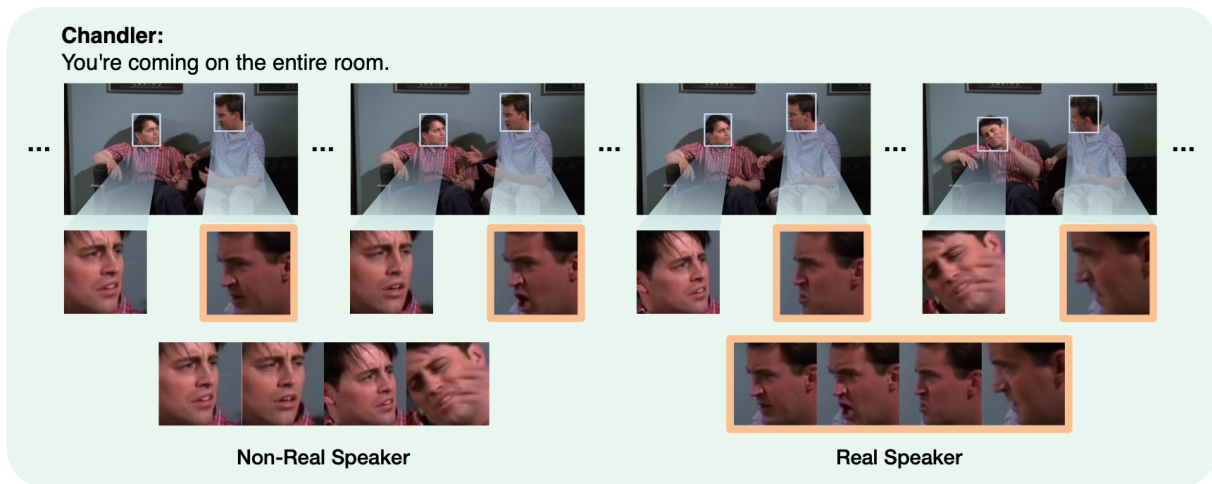


Figure 1: An example of MERMC task where an utterance contains two individuals with different facial expressions. One (*Joey* on the left side of the frame) expresses *disgust*, while the other (*Chandler* on the right side of the frame) expresses *anger*, and the latter is the real speaker whose emotion is annotated as the emotion of the utterance.

2022b) propose to detect all the faces in an utterance based on face detection tools such as MTCNN (Zhang et al., 2016), OpenFace (Baltrušaitis et al., 2016) or pre-trained active speaker detection models (Tao et al., 2021). However, given an utterance, the face sequence extracted by these methods may still contain multiple people, which may mislead the emotion prediction of the real speaker. For example, in Figure 1, there are two persons, *Joey* and *Chandler*, with distinct facial expressions, i.e., *disgust* and *anger*. Previous methods use the face sequence containing both persons’ faces as visual features, which will inevitably have a negative impact on predicting *Chandler’s* emotion. Therefore, to fully leverage the visual modality for emotion recognition, it is crucial to extract the face sequence of the real speaker of each utterance.

To this end, we propose a two-stage multimodal multi-task learning framework named **FacialMMT** for the MERMC task. In the first stage, we design a pipeline solution to obtain the face sequence of the real speaker, which contains three steps: 1) Extract the face sequence containing all possible speakers based on the combination of multimodal rules and an active speaker detection model (Tao et al., 2021); 2) Identify the number of face clusters in the face sequence based on an unsupervised clustering algorithm named InfoMap (Rosvall and Bergstrom, 2008); 3) Perform face matching and choose the face sequence with the highest confidence as the face sequence of the real speaker. Table 1 illustrates the differences between our method and previous

methods.

Based on the extracted face sequence, in the second stage, we further propose a **Multimodal facial expression-aware multi-task learning model** named MARIO. MARIO first resorts to an auxiliary frame-level facial expression recognition task to obtain the emotion distribution of each frame in the face sequence. Next, the emotion-aware visual representation is then integrated with textual and acoustic representations via Cross-Modal Transformer (Tsai et al., 2019) for utterance-level emotion recognition.

Our main contributions can be summarized as follows:

- To obtain the face sequence of the real speaker in an utterance, we propose a face sequence extraction method, which consists of three steps, i.e., multimodal face recognition, unsupervised face clustering, and face matching.
- We propose a **Multimodal facial expression-aware multi-task learning model** named MARIO for the MERMC task, which leverages an auxiliary frame-level facial expression recognition task to obtain the frame-level emotion distribution to help utterance-level emotion recognition.
- Experimental results on a benchmark dataset MELD demonstrate the superiority of our proposed FacialMMT framework over the SOTA systems. Moreover, FacialMMT outperforms a number of SOTA systems with a significant margin when only visual modality is used.

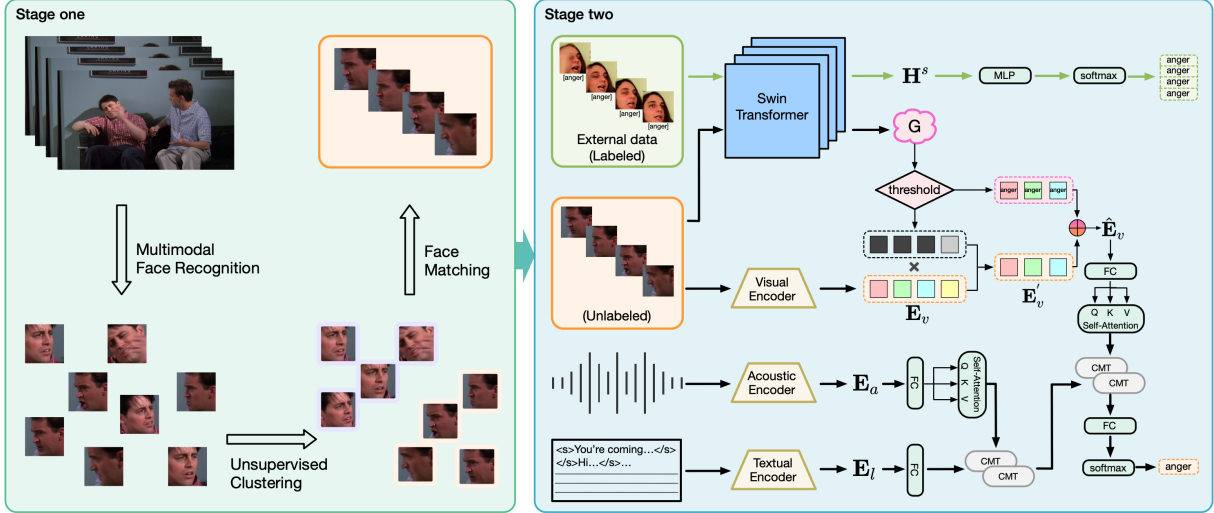


Figure 2: The overview of our FacialMMT framework. The first stage extracts the real speaker’s face sequence, and the second stage proposes a multimodal facial expression-aware multi-task learning model (MARIO) for MERMC.

2 Method

2.1 Task Formulation

Given an MERMC corpus \mathbb{D} , let us use $\{X_1, X_2, \dots, X_{|\mathbb{D}|}\}$ to denote a set of samples in the corpus. Each sample contains a multimodal dialogue with n utterances $d = \{u_1, u_2, \dots, u_n\}$, in which each utterance $u_i = \{u_{il}, u_{ia}, u_{iv}\}$ contains information from three modalities, i.e., text, audio, and vision, denoted by $\{l, a, v\}$. The goal of the MERMC task is to classify each utterance u_i into one of C pre-defined emotion types y_i , and predict a label sequence $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ for d . Note that each utterance is only annotated with one speaker’s identity (real speaker) and his/her emotion is annotated as the emotion of the current utterance.

2.2 Framework Overview

As shown in Figure 2, our FacialMMT framework contains two stages. To obtain the face sequence of the real speaker in each utterance, the first stage introduces a pipeline method to perform multimodal face recognition and unsupervised clustering, followed by face matching. With the extracted face sequences, the second stage resorts to an auxiliary frame-level facial expression recognition task to generate the emotion distribution for each frame in the face sequence, and then employs Cross-Modal Transformer to integrate the emotion-aware visual representations with text and acoustic representations for multimodal emotion recognition.

We will present the details of the two stages in the following two subsections.

2.3 Face Sequence Extraction

As shown in the left side of Figure 2, the first stage extracts the face sequence of the real speaker based on the following three steps:

Multimodal Face Recognition. First, we propose to combine multimodal rules and an active speaker detection (ASD) model to extract face sequences of all possible speakers. Specifically, given a video utterance, we use a pre-trained ASD model TalkNet (Tao et al., 2021) to combine visual and audio information for speaker detection. However, TalkNet often fails to identify speakers for videos with short duration or complex scenes (e.g., for a video with multiple people, someone is laughing or making noise instead of speaking). To obtain the face sequence in these videos, we further design several multimodal rules, including the opening and closing frequency of mouth, movement of different people’s mouths between video frames, and the alignment between the mouth movement and audio signals. The details of these multimodal rules are described in Appendix A.1.

Unsupervised Clustering. Based on the raw face sequence, we apply an unsupervised clustering algorithm InfoMap (Rosvall and Bergstrom, 2008) to identify the number of face clusters in the sequence as follows:

- We first employ the K-Nearest Neighbors algorithm to construct a graph of all potential speakers’ faces, and then calculate the similarity between faces, followed by using the normalized as the weight of edges.
- Random walks are then conducted on the graph

to generate different face sequences.

- Lastly, we hierarchically encode the face sequences, and minimize the minimum average encoding length to obtain the clustering result.

The minimization process includes minimizing the average encoding length of classes, as well as the average encoding length of each class’s in-class objects. The formulation is defined as follows:

$$\arg \min_{K, Y} \mathcal{L}(P, K, Y) = q_{i\curvearrowright} \left(- \sum_{i=1}^K \frac{q_{i\curvearrowright}}{q_{i\curvearrowright}} \log \frac{q_{i\curvearrowright}}{q_{i\curvearrowright}} \right) + \sum_{i=1}^K p_{i\circ} \left(- \frac{q_{i\curvearrowright}}{p_{i\circ}} \log \frac{q_{i\curvearrowright}}{p_{i\circ}} - \sum_{\alpha \in i} \frac{p_{\alpha}}{p_{i\circ}} \log \frac{p_{\alpha}}{p_{i\circ}} \right) \quad (1)$$

where Y is the predicted face sequence category, K represents the number of face sequences, $q_{i\curvearrowright}$ represents the probability of the occurrence of category i , $q_{i\curvearrowright} = \sum_{i=1}^K q_{i\curvearrowright}$, p_{α} represents the probability of the occurrence of a face image α , and $p_{i\circ} = q_{i\curvearrowright} + \sum_{\alpha \in i} p_{\alpha}$.

Face Matching. Finally, we construct a face library to determine the face sequence of the real speaker. Because the benchmark dataset for the MERMC task, i.e., MELD (Poria et al., 2019), contains six leading roles occurring frequently in the dataset, we manually select 20 different face images for each leading role based on the raw face sequence extracted in Multimodal Face Recognition and regard these 120 images as the face library. Next, we use a ResNet-50 model (He et al., 2016) pre-trained on a face recognition dataset MS-Celeb-1M (Guo et al., 2016) to extract visual features for the images in the library and in different face clusters. As each utterance provides the real speaker’s identity, who is either one of six leading roles or a passerby, we match the images in each face cluster with six leading roles’ images in the library by calculating the cosine similarity between their visual representations. Specifically, if the identity of the real speaker is one of the six leading roles, the face sequence with the highest similarity is regarded as the real speaker’s face sequence; otherwise, we regard the face sequence with the lowest similarity as the real speaker’s face sequence.

2.4 A Multimodal Facial Expression-Aware Multi-Task Learning Model

After obtaining the real speaker’s face sequence in each utterance, we further propose a Multimodal facial expression-aware multi-task learning model (MARIO), as shown in the right side of Figure 2.

Next, we introduce the details of MARIO, including unimodal feature extraction, emotion-aware visual representation, and multimodal fusion.

2.4.1 Unimodal Feature Extraction

In the MERMC task, given an utterance u_i , we extract unimodal features from three modalities $\{u_{il}, u_{ia}, u_{iv}\}$ to obtain the text, audio, and visual representations as follows:

- **Text:** To efficiently utilize the dialogue context and speaker’s emotional dynamics, we concatenate the input utterance and all its contextual utterances as input, and feed it into a pre-trained language model (e.g., BERT) for fine-tuning. We then take out the hidden representation of the first token as the text representation $\mathbf{E}_l \in \mathbb{R}^{d_l}$, where $d_l = 512$ is the size of text features.
- **Audio:** We obtain the word-level audio representation based on the Wav2vec2.0 model (Baevski et al., 2020) pre-trained on the Librispeech-960h dataset (Panayotov et al., 2015), denoted by $\mathbf{E}_a \in \mathbb{R}^{d_a}$, where $d_a = 768$ is the dimension of audio features.
- **Vision:** Given the real speaker’s face sequence of the input utterance, we use an InceptionResNetv1 model (Szegedy et al., 2017) pre-trained on the CASIA-WebFace dataset (Yi et al., 2014) to obtain the frame-level visual representation $\mathbf{E}_v \in \mathbb{R}^{L \times d_v}$, where L is the face sequence length and $d_v = 512$ is the size of visual features.

2.4.2 Emotion-Aware Visual Representation

Because the goal of MERMC is to predict the emotion of all the utterances in a dialogue, we propose to enhance the frame-level visual representation with the emotion distribution of each frame. To achieve this, we introduce an auxiliary frame-level facial expression recognition task, which is known as Dynamic Facial Expression Recognition (DFER) in the computer vision community (Li and Deng, 2020). Formally, let \mathbb{D}^s be another set of samples for the DFER task. Each sample is a face sequence containing m faces, denoted by $s = \{s_1, s_2, \dots, s_m\}$. The goal of DFER is to predict the label sequence $z = \{z_1, z_2, \dots, z_m\}$, where each label z_i belongs to one of C pre-defined facial expressions (i.e., emotion categories).

Auxiliary DFER Module. As shown in the top right of Figure 2, we employ a well-known Swin-Transformer model (Liu et al., 2021) pre-trained

on the Ms-Celeb-1M dataset (Guo et al., 2016) to obtain the representation of each frame in the face sequence as follows:

$$\mathbf{H}^s = \{\mathbf{h}_1^s, \dots, \mathbf{h}_m^s\} = \text{Swin-Transformer}(s) \quad (2)$$

where $\mathbf{H}^s \in \mathbb{R}^{m \times d_u}$ is the generated facial features. Next, we feed \mathbf{H}^s into a multi-layer perceptron (MLP) layer for facial expression recognition. During the training stage, we use cross-entropy loss to optimize the parameters for the DFER task:

$$p(z_i) = \text{softmax}(\text{MLP}(\mathbf{h}_i^s)) \quad (3)$$

$$\mathcal{L}^{DFER} = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^m \log p(z_{ij}) \quad (4)$$

where M is the number of samples in \mathbb{D}^s .

Facial Expression Perception for MERMC.

Based on the auxiliary DFER module, a direct solution to obtain the emotion-aware visual representation is to convert the predicted emotion of each frame to an one-hot vector and concatenate it with its original representation as the frame-level visual representation. However, as we all know, the one-hot vector derived from the *argmax* function is not differentiable, which will affect the parameter optimization in our multi-task learning framework.

To tackle this issue, we apply Gumbel Softmax Jang et al. (2017), which has a continuous relaxed categorical distribution, to obtain an approximated emotion distribution for each frame. By using softmax as the differentiable approximation of *argmax* and adding a temperature function τ , it can achieve gradient updates during backpropagation:

$$\mathbf{g}_i = \text{softmax}((g + \mathbf{h}_i^s)/\tau) \quad (5)$$

where $\mathbf{g}_i \in \mathbb{R}^C$, $g = -\log(-\log(u))$ is a noise sampled from the Gumbel distribution, and $u \sim \text{Uniform}(0, 1)$. As $\tau \rightarrow 0$, the softmax computation smoothly approaches the *argmax*, and the sample vectors approximate one-hot vectors.

Moreover, if the emotion distribution of the i -th frame in the face sequence concentrates on certain emotion, it shows that this frame reflects the clear emotion; otherwise if the emotion distribution is uniform, it implies the emotion in this frame is blurred and may bring noise to our MERMC task. To alleviate the noise from the emotion-blurred frames, we design a gating mechanism to dynamically control the contribution of each frame in the face sequence for the MERMC task. Specifically,

the emotion clarity of the i -th frame can be computed as $\delta_i = \mathbf{g}_i \cdot \mathbf{g}_i^\top$, where \cdot denotes the dot product. Based on this, we can obtain the emotion clarity of all the frames in the face sequence:

$$\delta = \{\delta_1, \delta_2, \dots, \delta_m\} \quad (6)$$

We then apply δ to the original visual representation \mathbf{E}_v to filter out the emotion-blurred frames, in which δ_i is less than a predetermined threshold.

Finally, we concatenate the filtered visual representation \mathbf{E}'_v and the emotion distributions of all the frames \mathbf{E}_e to obtain the emotion-aware visual representation as follows:

$$\hat{\mathbf{E}}_v = \mathbf{E}'_v \oplus \mathbf{E}_e, \quad \mathbf{E}_e = \{\mathbf{g}_1, \dots, \mathbf{g}_{m'}\} \quad (7)$$

where $\hat{\mathbf{E}}_v \in \mathbb{R}^{m' \times (d_v + C)}$, m' is the number of filtered frames, and \oplus is the concatenation operator.

2.4.3 Multimodal Fusion

Intra-Modal Interactions. We feed \mathbf{E}_a and $\hat{\mathbf{E}}_v$ to two separate self-attention Transformer layers (Vaswani et al., 2017) to model the intra-modal interactions within audio features and visual features as follows:

$$\mathbf{H}_a = \text{Transformer}(\mathbf{E}_a), \mathbf{H}_v = \text{Transformer}(\hat{\mathbf{E}}_v)$$

Inter-Modal Interactions. To achieve interactions between different modalities, we apply the Cross-Modal Transformer (CMT) layer (Tsai et al., 2019). Firstly, we fuse the text and audio modalities, alternating the two modalities as the query vector, then concatenating them to obtain the text-audio fused representation \mathbf{H}_{l-a} . Similarly, \mathbf{H}_{l-a} is then fused with visual modality to obtain the utterance-level text-audio-visual fused representation \mathbf{H}_{l-a-v} below:

$$\mathbf{H}_{l-a} = \text{CM-Transformer}(\mathbf{E}_l, \mathbf{H}_a) \quad (8)$$

$$\mathbf{H}_{l-a-v} = \text{CM-Transformer}(\mathbf{H}_{l-a}, \mathbf{H}_v) \quad (9)$$

Finally, \mathbf{H}_{l-a-v} is fed to a softmax layer for emotion classification:

$$q(y) = \text{softmax}(\mathbf{W}^\top \mathbf{H}_{l-a-v} + \mathbf{b}) \quad (10)$$

The standard cross-entropy loss is used to optimize the parameters for the MERMC task:

$$\mathcal{L}^{MERMC} = -\frac{1}{N} \sum_{i=1}^N \log q(y_i) \quad (11)$$

where N is the number of utterance samples.

The pseudocode for training the MARIO model is provided in Appendix A.2.

3 Experiments and Analysis

3.1 Dataset

To verify the effectiveness of our FacialMMT framework, we conduct experiments with two datasets. One is the dataset for the main MERMC task, and the other is the dataset for the auxiliary DFER task. The descriptions are as follows:

Dataset for MERMC: We use the MELD dataset (Poria et al., 2019), which is a publicly available dataset for MERMC. MELD contains 13,707 video clips extracted from the sitcom *Friends*, which contain information such as utterance, audio, video, and speaker identity. It also provides emotion annotations on each utterance with seven classes, including *neutral*, *surprise*, *fear*, *sadness*, *joy*, *disgust*, and *anger*.

Dataset for DFER: For the auxiliary DFER task, we use the Aff-Wild2 dataset (Kollias and Zafeiriou, 2019; Kollias, 2022), which contains 548 video clips collected from YouTube in real-world environments. Each clip has several frames of aligned faces and each frame is annotated with a facial expression. It has eight classes of emotions (six basic emotions, *neutral*, and *other*). Because the goal is to leverage Aff-Wild2 to guide the emotion prediction on our main dataset, we removed samples annotated with the *other* emotion.

3.2 Compared Systems

We compare FacialMMT against the following systems: **DialogueRNN** (Majumder et al., 2019) models the speaker identity, historical conversation, and emotions of previous utterances with RNNs. **ConGCN** (Zhang et al., 2019) proposes a Graph Convolutional Network (GCN)-based model, which constructs a heterogeneous graph based on context-sensitive and speaker-sensitive dependencies. **MMGCN** (Hu et al., 2021) builds both long-distance dependency and dependencies between speakers with GCNs. **DialogueTRM** (Hu et al., 2021) proposes to consider the temporal and spatial dependencies and models local and global context information. **DAG-ERC** (Shen et al., 2021) models the information flow between the conversation background and its surrounding context. **MM-DFN** (Hu et al., 2022a) introduces a dynamic fusion module to fuse multimodal context features. **EmoCaps** (Li et al., 2022b) extracts the emotional tendency and fuses modalities through an emotion capsule. **UniMSE** (Hu et al., 2022b) unifies multimodal sentiment analysis and ERC tasks with a

unified framework based on T5 (Raffel et al., 2020). **GA2MIF** (Li et al., 2023) proposes a graph and attention based two-stage multi-source multimodal fusion approach.

3.3 Implementation

For our FacialMMT framework, we employ either BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) as the textual encoder and use tiny version of Swin Transformer¹. The maximum length of input text is set to 512. The intercept operation is to remove the last word of the longest utterance in a dialogue and loop until the condition is met. The maximum length of visual and audio is set to the average plus 3 times the standard deviation. The batch sizes for MERMC task and DFER task is set to 1 and 150, respectively. The size of hidden layers is 768. The number of heads in self-attention layers and cross-modal transformer layers is 12, and the learning rates for MERMC and DFER is set to $7e-6$ and $5e-5$, respectively. The dropout rate is 0.1. The threshold for filter out the emotion-blurred frames is set to 0.2.

Following previous works, we use the weighted average F1-score as the evaluation metric for the MERMC task. For the DFER task, the macro F1-score on the validation set is reported. Our model is trained on a GeForce RTX 3090Ti GPU and parameters are optimized through an AdamW optimizer.

3.4 Main Results on the MERMC task

We report the results of different methods on the MERMC task in Table 2 and Table 4. The results of baselines are retrieved from previous studies.

First, we compare the multimodal emotion recognition results of each method. As shown in Table 2, **FacialMMT-RoBERTa** outperforms all the compared systems with a significant margin, indicating the effectiveness of our proposed approach. Additionally, we find that using BERT instead of RoBERTa as the text encoder leads to a slight decrease in performance. Although it performs relatively than the T5-based **UniMSE** model, it still outperforms all the other baseline systems that either use BERT or RoBERTa as the text encoder.

Moreover, we compare the emotion recognition results in a single visual modality. As shown in Table 4, previous methods such as **EmoCaps** and **MM-DFN** directly employ 3D-CNN to extract the visual features, which introduce environmental

¹<https://github.com/JDAI-CV/FaceX-Zoo>

Models	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	F1
<i>DialogueRNN</i> (Majumder et al., 2019)	73.50	49.40	1.20	23.80	50.70	1.70	41.50	57.03
ConGCN (Zhang et al., 2019)	76.70	50.30	8.70	28.50	53.10	10.60	46.80	59.40
MMGCN (Hu et al., 2021)	-	-	-	-	-	-	-	58.65
DialogueTRM* (Hu et al., 2021)	-	-	-	-	-	-	-	63.50
DAG-ERC* (Shen et al., 2021)	-	-	-	-	-	-	-	63.65
MM-DFN (Hu et al., 2022a)	77.76	50.69	-	22.94	54.78	-	47.82	59.46
EmoCaps* (Li et al., 2022b)	77.12	63.19	3.03	42.52	57.50	7.69	57.54	64.00
UniMSE [▲] (Hu et al., 2022b)	-	-	-	-	-	-	-	65.51
GA2MIF (Li et al., 2023)	76.92	49.08	-	27.18	51.87	-	48.52	58.94
FacialMMT-BERT	78.55	58.17	13.04	38.51	61.10	30.30	53.66	64.69
FacialMMT-RoBERTa	80.13	59.63	19.18	41.99	64.88	18.18	56.00	66.58

Table 2: Comparison results of the MERMC task on the MELD dataset. The baselines with italics only use textual modality. [▲] indicates the model uses T5 (Raffel et al., 2020) as the textual encoder. The baselines tagged with * and * respectively use BERT and RoBERTa as textual encoders. The best results are marked in bold.

	Savchenko (2022)	FacialMMT
F1	40.67	42.19

Table 3: Results of the DFER task based on F1 score.

noise and thus obtain relatively poor emotion recognition results. By extracting the face sequence of all possible speakers in a video, **MMGCN** achieves the best performance among the baseline systems. Moreover, we can observe our FacialMMT framework significantly outperforms all the compared systems, mainly due to the accurate extraction of the real speaker’s face sequence.

Lastly, we conduct ablation studies of Face Sequence Extraction in Section 2.3. First, after removing unsupervised clustering (UC) and face matching (FM), the emotion recognition result decreases by 2.12%, which demonstrates the usefulness of the two modules. Furthermore, if all the three steps are removed, meaning that video frames are directly used as visual features, the performance drops significantly.

3.5 Results on the DFER task

Table 3 shows the comparison of our method and one of the state-of-the-art methods (Savchenko, 2022) on the DFER task. For a fair comparison, we re-implement the compared system and run experiments based on the same setting as ours. In Table 3, we can clearly observe that our framework outperforms the compared system by 1.52 absolute percentage points on the Aff-Wild2 dataset, which demonstrates the effectiveness of our model on the auxiliary DFER task.

Models	Composition of visual information	F1
EmoCaps	Video frames	31.26
MM-DFN	Video frames	32.34
MMGCN	Possible speakers’ face sequences	33.27
	Real speaker’s face sequence	36.48
FacialMMT	- w/o UC, FM	34.36
	- w/o MFR, UC, FM	32.27

Table 4: Comparison of single visual modality emotion recognition results. MFR represents multimodal face recognition, UC represents unsupervised clustering, and FM represents face matching.

3.6 Ablation Study

We conduct ablation studies of FacialMMT, and show the results in Table 5. It is obvious that any removal of one or two modality leads to a performance drop, indicating that any modality plays an essential role in emotion recognition. Specifically, we can infer that the visual modality plays a more important role than the audio modality, which differs from the observations in previous studies. This suggests that enhancing multimodal emotion recognition from the perspective of visual representation is effective. Moreover, removing the auxiliary DFER module also drops the performance, indicating that introducing frame-level facial expression supervision signals can indeed provide important clues for utterance-level emotion recognition.

3.7 Case Study

To better understand the two main contributions of our work, we present two examples in Figure 3. As our work is the first to use enhanced visual representations to help the MERMC task, we only compare FacialMMT with its variants: 1) **Toolkit-based FacialMMT** represents using face detection



Figure 3: Prediction comparison between different methods on two test samples for the MERMC task.

FacialMMT	66.58
- w/o Audio	66.20
- w/o Vision	65.55
- w/o Audio, Vision	63.98
- w/o Text, Vision	38.02
- w/o Auxiliary DFER Module	66.08

Table 5: Ablation study of FacialMMT based on F1 score.

toolkits to detect the face sequence, as in previous methods; 2) **FacialMMT -w/o UC, FM** represents face sequences extracted by multimodal face recognition. As shown in Figure 3, the face sequences extracted by two variants of our model contain much noise, which may mislead the emotion prediction of the input utterance. In contrast, our **FacialMMT** model correctly extracts the face sequence of the real speaker in both cases, and leverages the frame-level emotion distribution to help correctly predict the utterance-level emotions.

4 Related Work

4.1 Emotion Recognition in Conversations

Recently, Emotion Recognition in Conversations (ERC) has gradually become a hot topic in the field of emotion analysis. According to the input form, ERC is classified into text-based ERC and multimodal ERC. Text-based ERC mainly focuses on research in modeling context, modeling speaker relationships, and incorporating common-sense knowledge (Majumder et al., 2019; Li et al., 2020; Shen et al., 2021; Liu et al., 2022c; Li et al., 2022a; Ong et al., 2022).

To better mimic the way of human thinking, mul-

timodal ERC has rapidly developed in recent years. Multimodal ERC mainly focuses on multimodal feature extraction, interaction, and fusion. First, some studies (Mao et al., 2021; Joshi et al., 2022; Li et al., 2022a) consider context information in conversations and utilize pre-trained language models such as BERT (Devlin et al., 2019) and BART (Lewis et al., 2020) to obtain dialogue-level text representations. Some works (Dai et al., 2021; Liang et al., 2021; Hu et al., 2021; Zhao et al., 2022b) also extract facial representations using various tools, such as MTCNN (Zhang et al., 2016). For multimodal interactions, exiting studies (Tsai et al., 2019; Lv et al., 2021) propose a Cross-Modal Transformer model and a progressive modality reinforcement approach for unaligned multimodal sequences. For modality fusion, Jin et al. (2020) propose a localness and speaker aware transformer to capture local context and emotional inertia. Li et al. (2022b) design an emotion capsule to fuse sentence vectors through multimodal representations, and Zou et al. (2022) propose to use a main modal Transformer to improve the effectiveness of multimodal fusion. In this work, due to the specific nature of multi-party conversations, we extract the face sequence of the real speaker from a video, and use frame-level facial expressions to help utterance-level emotion recognition.

4.2 Dynamic Facial Expression Recognition

The value of understanding facial expressions lies in collecting direct impressions from others during a conversation. Thus, there has been a significant amount of research conducted on the Dynamic Fa-

cial Expression Recognition (DFER) task. Early DFER datasets were mainly collected from laboratory environments, such as CK+ (Lucey et al., 2010), MMI (Valstar et al., 2010), Oulu-CASIA (Zhao et al., 2011). Since 2013, Emotion Recognition in the Wild (EmotiW) competition has been held, researchers have begun to shift their focus from laboratory-controlled environments to more realistic and complex wild scenarios. Some works (Sümer et al., 2021; Delgado et al., 2021; Mehta et al., 2022) focus on predicting student engagement, while others focus on mental health issues (Yoon et al., 2022; Amiriparian et al., 2022; Liu et al., 2022a). Moreover, there are also several studies proposing new datasets or methods for facial expression recognition of characters in movies and TV shows (Jiang et al., 2020; Zhao and Liu, 2021; Toisoul et al., 2021; Wang et al., 2022b; Liu et al., 2022b).

5 Conclusion

In this paper, we proposed a two-stage framework named **Facial** expression-aware **Multimodal Multi-Task** learning (FacialMMT) for the MERMC task. FacialMMT first extracts the real speaker’s face sequence from the video, and then leverages an auxiliary frame-level facial expression recognition task to obtain the emotion-aware visual representation through multi-task learning, followed by multimodal fusion for the MERMC task. Experiments on the MELD dataset show the effectiveness of FacialMMT.

Limitations

Our work has the following limitations. First, our proposed FacialMMT approach is a two-stage framework that is not fully end-to-end. We plan to propose an end-to-end framework in the future, which integrates face sequence extraction and multimodal emotion recognition in a joint learning manner. Second, this work primarily focuses on the visual modality, and has not yet delved into other aspects of the MERMC task. Therefore, in the future, we plan to leverage the extracted face sequences to explore better cross-modal alignment and multimodal fusion mechanisms to improve the performance of the MERMC task.

Ethics Statement

We would like to thank Poria et al. (2019) and Kollias and Zafeiriou (2019) for their valuable work

in constructing and sharing the MELD and Aff-Wild2 datasets. MELD is licensed under the GNU General Public License v3.0². For the Aff-Wild2, we have signed an End User License Agreement³. Since MELD is built on the sitcom Friends, we manually annotate 20 different face images occurring in the sitcom for each of the six leading roles without accessing to any personal information. We do not share personal information and do not release sensitive content that can be harmful to any individual or community.

If applying our framework to real-world scenarios in the future, it could potentially involve some ethical issues such as user privacy and ethical biases, as pointed out by Stark and Hoey (2021) and Stark and Hutson (2021). While the ethical issues faced in emotion recognition are common, we will engage with the concerns raised about emotion recognition in the references and strictly comply with relevant regulations and ethical standards. Specifically, our work is based on publicly available datasets, and if we construct new MERMC datasets in the future, we will carefully consider user privacy issues, anonymize or obfuscate facial data, and ensure that the framework is only used in contexts where explicit consent for facial data processing has been obtained. Moreover, we will refer to the recommendations in Stark and Hoey (2021) and develop a comprehensive ethical framework that guides our research process. We are also committed to being transparent about our research methods, data sources, and potential limitations. Regarding potential biases, we plan to evaluate our framework on more diverse datasets in the future, and propose appropriate solutions to alleviate the bias issues.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This work was supported by the Natural Science Foundation of China (62076133 and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (BK20200463) and Distinguished Young Scholars (BK20200018).

²<https://github.com/declare-lab/MELD/blob/master/LICENSE>

³<https://ibug.doc.ic.ac.uk/resources/cvpr-2022-3rd-abaw/>

References

- Shahin Amiriparian, Lukas Christ, Andreas König, Eva-Maria Meßner, Alan Cowen, Erik Cambria, and Björn W Schuller. 2022. Muse 2022 challenge: Multimodal humour, emotional reactions, and stress. In *Proceedings of ACM MM*.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of NeurIPS*.
- Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Proceedings of WACV*.
- 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*.
- Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of NAACL-HLT*.
- Kevin Delgado, Juan Manuel Origgi, Tania Hasanpoor, Hao Yu, Danielle Allessio, Ivon Arroyo, William Lee, Margrit Betke, Beverly Woolf, and Sarah Adel Bargal. 2021. Student engagement dataset. In *Proceedings of ICCV*.
- 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*.
- Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of ECCV*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*.
- Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022a. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *Proceedings of ICASSP*.
- Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of EMNLP*.
- 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*.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparametrization with gumble-softmax. In *Proceedings of ICLR*.
- Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2010. 3d convolutional neural networks for human action recognition. *Proceedings of ICML*.
- Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuang-gao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. Dfaw: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of ACM MM*.
- Xiao Jin, Jianfei Yu, Zixiang Ding, Rui Xia, Xiangsheng Zhou, and Yaofeng Tu. 2020. Hierarchical multimodal transformer with localness and speaker aware attention for emotion recognition in conversations. In *Proceedings of NLPCC*.
- Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. 2022. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of NAACL-HLT*.
- Dimitrios Kollias. 2022. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of CVPR*.
- Dimitrios Kollias and Stefanos Zafeiriou. 2019. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*.
- 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*.
- Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. Ga2mif: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Trans. Affect. Comput.*
- Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of COLING*.
- Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.*
- Shimin Li, Hang Yan, and Xipeng Qiu. 2022a. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of AAAI*.
- Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022b. Emocaps: Emotion capsule based model for conversational emotion recognition. In *Proceedings of ACL (Findings)*.
- Jingjun Liang, Ruichen Li, and Qin Jin. 2020. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of ACM MM*.

- Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In *Proceedings of AAAI*.
- Feng Liu, Han-Yang Wang, Si-Yuan Shen, Xun Jia, Jing-Yi Hu, Jia-Hao Zhang, Xi-Yi Wang, Ying Lei, Ai-Min Zhou, Jia-Yin Qi, et al. 2022a. Opo-fcm: A computational affection based occ-pad-ocean federation cognitive modeling approach. *IEEE Trans. Comput. Soc. Syst.*
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022b. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of ACM MM*.
- Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. 2022c. Dialogueein: Emotion interaction network for dialogue affective analysis. In *Proceedings of COLING*.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of ICCV*.
- Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of CVPR*.
- Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of CVPR*.
- Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of ACL*.
- Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI*.
- Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. Dialoguetrm: Exploring multimodal emotional dynamics in a conversation. In *Proceedings of EMNLP (Findings)*.
- Naval Kishore Mehta, Shyam Sunder Prasad, Sumeet Saurav, Ravi Saini, and Sanjay Singh. 2022. Three-dimensional densenet self-attention neural network for automatic detection of student’s engagement. *Appl. Intell.*
- Donovan Ong, Jian Su, Bin Chen, Anh Tuan Luu, Ashok Narendranath, Yue Li, Shuqi Sun, Yingzhan Lin, and Haifeng Wang. 2022. Is discourse role important for emotion recognition in conversation? In *Proceedings of AAAI*.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Proceedings of ICASSP*.
- 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*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*
- Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*.
- Andrey V Savchenko. 2022. Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In *Proceedings of CVPR*.
- Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of ACL*.
- Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of ACM FAccT*.
- Luke Stark and Jevan Hutson. 2021. Physiognomic artificial intelligence. *Fordham Intell. Prop. Media & Ent. LJ*.
- Ömer Sümer, Patricia Goldberg, Sidney D’Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2021. Multimodal engagement analysis from facial videos in the classroom. *IEEE Trans. Affect. Comput.*
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of AAAI*.
- Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of ACM MM*.

- Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.*
- Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of ICCV*.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*.
- Michel Valstar, Maja Pantic, et al. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of LREC*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of NeurIPS*.
- Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2022a. Multimodal emotion-cause pair extraction in conversations. *IEEE Trans. Affect. Comput.*
- Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. 2022b. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of CVPR*.
- Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.
- Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of AAAI*.
- Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of IJCAI*.
- Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*
- Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen. 2011. Facial expression recognition from near-infrared videos. *Image Vis. Comput.*
- Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. 2022a. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP*.
- Jinming Zhao, Tengan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022b. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of ACL*.
- Zengqun Zhao and Qingshan Liu. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of ACM MM*.
- ShiHao Zou, Xianying Huang, XuDong Shen, and Han-kai Liu. 2022. Improving multimodal fusion with main modal transformer for emotion recognition in conversation. *Knowl. Based Syst.*

A Appendix

A.1 Multimodal Rules

We design several multimodal rules to obtain possible speakers' face sequences from a video. The detailed steps are as follows: 1) using the FFmpeg⁴ tool to sample frame-level images from a video; 2) using the OpenFace⁵ library to detect all the people in the frame-level images, and obtain different FaceID, confidence, 68 feature landmarks, and aligned facial images; 3) using the FFmpeg tool to extract the audio from the video; 4) determining the number of possible speakers in the current video based on three rules as follows:

- “Mouth open-close” count. For different FaceID candidates, count their mouth open and close times respectively. If the sum of the distance between the upper and lower lips of a person is greater than a certain threshold at a certain time, we will record that the mouth of this FaceID is open at the current time.
- Mouth movement. Count which person's mouth moves the most during the time period by considering the movement of the lips between two consecutive frames of facial images, the difference in width between the inner corners of the mouth between these two frames, and the difference in height between the upper and lower inner lips between these two frames.
- Voice Activity Detection algorithm. Following Zhao et al. (2022a), we identify which frames in the current video have sound by considering whether the visual movement of the lips matches the audio signal. The better the matching is, the higher the probability that it is the speaker.

⁴<https://ffmpeg.org/>

⁵<https://github.com/TadasBaltrusaitis/OpenFace>

A.2 Pseudo-code of MARIO

We provide the pseudocode for training the proposed MARIO model, where θ_{Swin} , θ_T , $\theta_{self-attn}$, and θ_{CMT} represent the parameters of Swin-Transformer, the text encoder, self-attention Transformer, and Cross-Modal Transformer, respectively.

Algorithm 1: Multitask training procedure of MARIO

Input: DFER dataset; MERMC dataset.

Output: θ_{Swin} ; θ_T ; $\theta_{self-attn}$; θ_{CMT} .

```

1 repeat
2   for all batches in the DFER dataset do
3     Forward face sequences through
4     Swin-Transformer ;
5     Compute loss  $\mathcal{L}^{DFER}$  ;
6     Finetune  $\theta_{Swin}$  using  $\nabla_{\mathcal{L}^{DFER}}$ 
7   for all batches in the MERMC dataset
8   do
9     Forward text through text encoder ;
10    Forward face sequences through
11    Swin-Transformer ;
12    Obtain facial expression-aware
13    visual representation ;
14    Audio and vision are sent to
15    self-attention Transformer layer
16    respectively ;
17    Conduct cross-modal fusion of text
18    and audio ;
19    Conduct cross-modal fusion of
20    text-audio and vision ;
21    Compute loss  $\mathcal{L}^{MERMC}$  ;
22    Update  $\theta_{self-attn}$  and  $\theta_{CMT}$  and
23    finetune  $\theta_{Swin}$  and  $\theta_T$  using
24     $\nabla_{\mathcal{L}^{MERMC}}$  ;
25 until epoch reaches its maximum;

```

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section Limitations
- A2. Did you discuss any potential risks of your work?
Section Ethics Statement
- A3. Do the abstract and introduction summarize the paper’s main claims?
Section Abstract and Section Introduction
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

We use several pre-training language models, which are referenced and briefly introduced in Section Method

- B1. Did you cite the creators of artifacts you used?
Section Method
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section Ethics Statement
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
For the scientific artifacts we use, we use them as expected; The scientific artifact we have created is described in the Section Introduction and its intended use.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section Ethics Statement
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We analyze our FacialMMT framework in Section Experiments and Analysis.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We show the details of the datasets we used in section Experiments and Analysis.

C Did you run computational experiments?

Section Experiments and Analysis.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We describe them in the experimental setting of Section 3.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We describe them in the experimental setting of Section 3.
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We describe them in Section 3.
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
we report them in Section 3.
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Section Method
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We ourselves annotated 20 different face images for each of the six leading occurring in the sitcom
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We ourselves annotated 20 different face images for each of the six leading occurring in the sitcom
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The data we use is publicly available
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We use publicly available datasets and have added a response to ethics review in the camare-ready submission and look forward to it being approved.
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We ourselves annotated data