AM^2-EMOJE: ADAPTIVE MISSING-MODALITY EMOTION RECOGNITION IN CONVERSATION VIA JOINT EMBEDDING LEARNING

Naresh Kumar Devulapally, Sidharth Anand, Sreyasee Das Bhattacharjee, Junsong Yuan
The State University of New York at Buffalo

ABSTRACT

Human emotion can be presented in different modes i.e., audio, video, and text. However, the contribution of each mode in exhibiting each emotion is not uniform. Furthermore, the availability of complete mode-specific details may not always be guaranteed in the test time. In this work, we propose AM^2-EmoJE, a model for Adaptive Missing-Modality Emotion Recognition in Conversation via Joint Embedding Learning model that is grounded on two-fold contributions: **First**, a *query adaptive fusion* that can automatically learn the relative importance of its mode-specific representations in a query-specific manner. By this the model aims to prioritize the mode-invariant spatial (within utterance) query details of the emotion patterns, while also retaining its mode-exclusive aspects within the learned multimodal query descriptor. **Second** the *multimodal joint embedding learning* module that explicitly addresses various missing modality scenarios in test-time. By this, the model learns to emphasize on the correlated patterns across modalities, which may help align the cross-attended mode-specific descriptors pairwise within a joint-embedding space and thereby compensate for missing modalities during inference. By leveraging the spatio-temporal details at the dialogue level, the proposed AM^2-EmoJE not only demonstrates superior performance (around $2 - 4\%$ improvement in Weighted-F1 scores) compared to the best-performing state-of-the-art multimodal methods, by effectively leveraging body language in place of face expression, it also exhibits an enhanced privacy feature. By reporting around $2-5\%$ improvement in the weighted-F1 score, the proposed multimodal joint embedding module facilitates an impressive performance gain in a variety of missing-modality query scenarios during test time.

Index Terms— multimodal emotion recognition, joint embedding learning, adaptive fusion, missing modality

1. INTRODUCTION

Emotion analysis and its dynamic evolution in humans during conversations is pivotal to various critical tasks ranging from sentiment analysis in social media to affect-aware human-robot interactions. The problem is particularly complex due to the involvement of multi-party stakeholders, their mutual interactions described via different modalities like text, video, and audio, as well as their body language, which influence
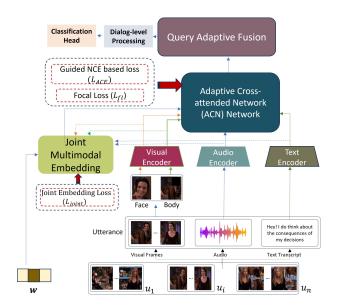

Fig. 1. Proposed AM^2-*EmoJE* framework

the speaker's emotions and its spatio-temporal progression expressed in an utterance.

In particular, a speaker's emotion depends on various intra (e.g. personal details, behavioral patterns, and habits) and inter (e.g. audience behavior, their interpreting conducts) personal contexts and other environmental circumstances. At one end, most multimodal methods assume data completeness, which may not hold in practice due to privacy, device, or security constraints. We note that the mode-specific descriptor components exhibiting an individual's emotion have some correlation, which may be instrumental to handling the missing modality information. However, this aspect has not yet been sufficiently explored by the existing literature. At the other end, the significance of this variety of heterogeneous mode information may not be uniform in evaluating emotions for all speakers. For example, a speaker's facial expression may have only limited information about their true emotion, while their body language may still be more semantically rich. Thus the challenges related to *the presence of strong heterogeneity both within and across multiple modalities and complex interplay* and *the influence of a missing modality information toward learning comprehensive multimodal interaction patterns in a real-life conversation setting* are critical yet under-studied.

Toward these, we propose AM^2-*EmoJE*, an *Adaptive*

*M*issing-*M*odality *Emo*tion Recognition in Conversation via *J*oint *E*mbedding Learning model that is grounded on the following contributions.

1. *Query Adaptive Fusion (QAF)* that derives the relative importance of the cross-attended mode descriptors on the fly to deliver a robust multimodal query descriptor preserving both instance-specific and category-specific utterance-level spatial patterns in parallel.

2. *Multimodal joint embedding* that may leverage an auxiliary boolean mask vector as user input to turn on an effective mode switching mechanism, which proves instrumental in delivering a competitive decision ability for queries with incomplete mode information.

3. *Extensive Evaluation Analysis* using publicly available (MELD [1], IEMOCAP [2]) datasets not only demonstrate an impressive classification performance ($2 - 4\%$ improvement in Weighted-F1 of AM^2-*EmoJE* in complete multimodal test scenarios, but also exhibit a competitive decision making for queries with missing modality information.

2. RELATED WORK

Multimodal solutions toward motion Recognition in Conversations (ERC) have recently demonstrated significant performance improvements compared to their unimodal counterparts. While natural language transcriptions serve as strong emotion indicators for unimodal ER pipelines, leveraging the advantages of cross-modal interactions demonstrates significant performance gain[3, 4, 5, 6].

While exhibiting promising performances, traditional multimodal methods assume data completeness, which may not be a feasible constraint for various practice issues related to privacy, device, or security constraints and missing modality in a test environment appears to be challenging to these models. In a recent work, Ma et al. [7] investigate the behavior of Transformer-based multimodal frameworks in the presence of modal-incomplete data. Lee et al. [8] propose a prompt learning framework to address the issues of missing modalities during training or testing. However, the limitation of computational resources still persists as a bottleneck to such training-heavy transformer models.

While a few recent works [9, 10, 11, 12] attempt to address the missing modality challenge in specific application settings, a straightforward extension to these preliminary models for the complex task of ERC, may not be feasible. Evaluating the relative contribution of multiple mode-specific components describing a speaker's emotion dynamics on the fly, while estimating their missing-mode description in parallel, is challenging. Toward these, we propose AM^2-*EmoJE* that is motivated by two-fold research objectives:**First**, a *query adaptive fusion* that can automatically learn the relative importance of its mode-specific representations in a query-specific manner; **Second** *multimodal joint embedding learning* framework that learns a mode-switching

mechanism to align the cross-attended mode-specific descriptors pairwise within a joint-embedding space and thereby allowing the model to compensate for missing modalities during inference.

3. PROPOSED METHOD

Problem Definition: Given a multi-party dialogue $d := \{u_j\}_j \in \mathcal{D}$ represented as a sequence of utterances $\{u_j\}_j$, the objective is to evaluate the dominant emotional states of the speaker as expressed in each utterance u_j. For brevity, now onward we will omit the suffix j, and an arbitrary utterance u_j will be represented as u unless the suffix is specifically required. Each $u \in \mathcal{D}$ contains (u_v, u_a, u_t), where u_v is the *video*, u_a is the *audio*, and u_t is the *text transcription* component of the utterance. In this work, we propose a continual emotion evaluation model that can evaluate the series of dialogues by single or multiple speakers in a conversational environment. Unlike existing multimodal literature, which primarily relies on the availability of complete mode-specific components representing a multimode utterance, in this work we allow incomplete query videos, i.e., some of the mode-specific information related to the query may be missing. Therefore, the availability information of each mode-specific query component is passed by a *boolean mask vector* $\mathbf{w} := (w_v, w_a, w_t)$. For example, a query that only has an audio and transcript will is passed with a mask vector $\mathbf{w}$ with $w_v = 0$. In another query environment, if a user is only interested in analyzing the visual component of a multimode query, then the mask vector $\mathbf{w}$ with $w_a = 0$ and $w_t = 0$ is passed with the query that may enable the system to perform only the visual analysis.

3.1. Mode-Specific Feature Representation
To capture the spatio-temporal evolution of information within each utterance, the first-level mode-specific feature representation scheme as described below:
3.1.1. Text Representation
To derive a compact descriptor for the text component u_t represented as a sequence of w words, i.e. $u_t = \{\omega_1, \omega_2, ..., \omega_w\}$, we employ the pretrained model SBERT [13] to obtain the fixed language embedding $\mathbf{f}^T \in \mathbb{R}^{w \times d_t}$ for the text component u_t.

3.1.2. Video Representation
For the visual component u_v of each utterance $u \in d$, FFmpeg is used to identify n key frames and MTCNN [**?**] is applied to extract the aligned faces from each key frame. To deliver a comprehensive analysis of the subjects' emotional evolution to appear in an utterance, we analyze their facial expression in contrast to their body language, each u_v is split into two sequences, which cover the visuals of distinct regions in the utterance: 'face sequence' that contains the sequence of derived frames containing only the subjects' face regions; 'body sequence' that exclusively contains the subjects' body segment in the utterance. YOLOv7 [14] fine-tuned for human detection, is employed to localize the human body segment in a

frame, To analyze the body language in exclusion, we subtract the face region from the identified body segment in a frame to design the derived frames in the 'body sequence'. Thus, the visual content u_v of u is represented in terms of two equal-sized derived frame sequences: $\mathbf{v}^{face} = \{\mathbf{au}_1, \mathbf{au}_2, ..., \mathbf{au}_n\}$ and $\mathbf{v}^{body} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n\}$, where each $\mathbf{au}_j$ and $\mathbf{b}_j$ represent a learned descriptor describing the jth element in $\mathbf{v}^{face}$ and $\mathbf{v}^{body}$ respectively. Two identical Bi-LSTM-based sequence representation modules with a hidden embedding size of d_v, which take $\mathbf{v}^{face}$ or $\mathbf{v}^{body}$ as inputs, are employed to obtain the initial regional descriptors $\mathbf{f}^{face} \in \mathbb{R}^{n \times d_v}$ or $\mathbf{f}^{body} \in \mathbb{R}^{n \times d_v}$. These two independent LSTM networks are then merged via a stacked self-attention layer that attends to the pair of these network-specific inputs to derive a self-attended visual descriptor $\mathbf{f}^v \in \mathbb{R}^{2n \times d_v}$ for u_v.

3.1.3. Audio Representation

We use the Pathout Fast 2D Spectrogram model (PASST) [15], which is initialized from a vision transformer trained on ImageNet, to fine-tuned on 10s audio clips from AudioSet. To derive a fine-grained understanding of the audio signal evolution, the audio component of each utterance is divided into smaller e overlapped segments and each segment is represented using their corresponding PASST descriptor $\mathbf{a}_i \in \mathbb{R}^{d_{passt}}$ such that $u_a :== \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_e\}$. Similar to the video representation, we use a BiLSTM to derive an overall audio feature for the utterance $\mathbf{f}^a \in \mathbb{R}^{d_a}$.

3.2. Weighted Multimodal Attention

Toward integrating the mode-specific contents across modalities, we note that the expression captured by these uni-mode components may not solely reflect their intra-mode data pattern. They may also represent the influence of information depicted by other modalities. For example, a speaker's audio may reflect certain emotions, which may not be sufficiently supported by their body language or an utterance transcript may be continually modulated by the expressions shown by the audience in the previous utterance segments. Furthermore, the availability of a quality mode-specific representation of an utterance may also not be guaranteed. For instance, the availability of visual cues may not be assured in every query scenario. Otherwise, noisy audio may also deteriorate the overall data quality occasionally.

While existing multimodal emotion analysis models do not allow such incomplete query information, we design an Adaptive Cross-attended Network (ACN) that enables a flexible yet robust feature representation technique to capture the effects of same- and different-modality cues at various levels of details (including missing modality cues). As observed in Figure 1, each layer of *ACN* is administered by multiple cross-modal cues from multiple *outer* networks into a single mode-specific *inner* network. The merging layer of *ACN* incorporates a simple yet effective masking mechanism toward facilitating the learning of several variants of mode combinations, wherein the model is explicitly trained

to control the contribution of each 'outer' network including a variety of missing-modality scenarios. The keys and values are generated as $\mathbf{K}_{m_i}^l = linear(w_{m_i}(W_{m_i}^{l,K})^T \mathbf{e}_{m_i})$, $\mathbf{V}_{m_i}^l = linear(w_{m_i}(W_{m_i}^{l,V})^T \mathbf{e}_{m_i})$, where $W^{l,K} \in \mathbb{R}^{d_{m_i} \times d_l}$ and $W^{l,V} \in \mathbb{R}^{d_{m_i} \times d_l}$ are key and value weight matrices in the lth layer of the attention network. The output of each attention head in the lth layer is then computed as:

$$\mathbf{g}_m^l = \mathbf{g}_m^{l-1} + \frac{1}{|\mathcal{M}|} \text{softmax}(linear(\frac{\mathbf{g}_m^{l-1}(\sum_{m_i \in \mathcal{M} \setminus \{m\}} \mathbf{K}_{m_i}^{l-1})^T}{\sqrt{d_{m_i}}}))\mathbf{V}_{m_i}^{l-1}$$
(1)

Thus, for each modality m in $\mathcal{M}$, the set of all modalities leveraged in the model, the learned descriptor $\mathbf{g}^m$ from multiple heads are mean-pooled to capture the weighted aggregated details within the final output of the *Central* network.

As shown in Figure 1, each mode-specific *Central* query network for each $m \in \mathcal{M}$ of *ACN* thus produces an average pooled cross-attended mode-specific descriptor $\mathbf{f}_{ACN}^m \in \mathbb{R}^d$ for the uni-mode components t, v, and a for u. We will discuss the learning algorithm later in Section 3.5.

3.3. Query Adaptive Fusion

In contrast to the existing fusion techniques, which assume that the mode-specific contributions in the resulting multimodal descriptor for a query input should be uniform, in reality, the quality and availability of the complete mode-specific components may not be feasible in general. Therefore, we propose a *Query Adaptive Fusion* (*QAF*) mechanism that designs a linear combination of the learned cross-attended mode descriptors $\mathbf{g}^m$ to define a comprehensive multimode utterance-level descriptor as follows:

$$\mathcal{A}(u) = \sum_{m \in \mathcal{M}} \left(\frac{w_m}{|\mathcal{M}|} \sum_{m_i \in \mathcal{M} \setminus m} \alpha_{m_i}^m \mathbf{f}^{m_i} + (1 - \alpha_{m_i}^m)\mathbf{f}^{m_i} \right)$$
(2)

where $0 \leq \alpha_{m_i}^m \ \forall m, m_i \in \mathcal{M} \leq 1$ are learnable parameters.

Thus, the proposed fusion function $\mathcal{A}$ provides a flexible multimodal representation mechanism, by which the resulting multimodal descriptor $\mathcal{A}(u)$ for an utterance u can retain category-specific discriminative data patterns, however not completely disregarding the unique instance-specific data patterns observed in the utterance. Furthermore, given an utterance u, in the absence of its complete mode-specific details (as provided by the user-specified mask vector $\mathbf{w}$) related to m, the model automatically learns to adopt its learnable parameters $\{\alpha_{m_i}^m\}_{m_i}$ to derive an adjusted multimodal descriptor $\mathcal{A}(u)$. We adopt the optimization approach of [16] to learn the interpolation parameters $\alpha_{m_i}^m$.

3.4. Classification

Given a conversational dialogue represented using a sequence of n utterances $\{u_j\}_j^n \in \mathcal{D}$, the emotion of the speaker s is estimated by leveraging two parallel utterance sequences: *Dialogue Context* that describes the sequence $\{\mathcal{A}(u_j)\}_j$; *Speaker Context* that describes a sub-sequence $\{\mathcal{A}(u_{s_j})\}_j$, where the sub-sequence $\{u_{s_j}\}_{s_j \in [1,n]}$ is generated from the dialogue

and includes only those utterances, in which s vocally contributes to the conversation. Two parallel Bi-LSTMs are trained to capture the spatio-temporal contexts independently of these contexts' perspectives: $\mathbf{s}_l \in \mathbb{R}^s$ representing the *Speaker Context* and $\mathbf{d}_l \in \mathbb{R}^s$ representing the *Dialogue Context*. These representations are passed through a classification head to arrive at the classification decision for the emotion for the utterance u.

3.5. Training objectives

The learning of AM^2-*EmoJE* includes two independent learning objectives: a multi-component loss objective ($\mathcal{L}$) that combinedly optimizes the guided NCE loss ($\mathcal{L}_{ACE}$) and the focal loss ($\mathcal{L}_{fl}$) for the classifier; a CLIP-like loss function $\mathcal{L}_{joint}$ to learn the pairwise joint embedding spaces. We discuss them below.

3.5.1. Guided NCE

The proposed Adaptive Cross-attended Network (*ACN*) jointly learns the cross-attended representations $\mathbf{f}_{ACN,j}^t$, $\mathbf{f}_{ACN,j}^v$, and $\mathbf{f}_{ACN,j}^a$ with twofold contributions: 1) preserving instance specific discriminability by identifying more reliable modes on the fly. We intuitively expect a speaker's face and body language to display similar emotions, and a significant deviation from this would require us to rely more on the other modalities like audio or transcript to accurately estimate their true emotions. Guided by this observation, we leverage an aggregated noise contrastive estimation ($\mathcal{L}_{ACE}$) as below:

$$\mathcal{L}_{ACE} = \frac{1}{|\mathcal{D}|} \sum_{u_j \in \mathcal{D}} \frac{1}{|\mathcal{M}|} \sum_{\substack{m \neq m_i \\ m, m_i \in \mathcal{M}}} \mathcal{L}_{\text{NCE}}(\mathbf{f}_{ACN,j}^m, \mathbf{f}_{ACN,j}^{m_i})$$

(3)

with

$$\mathcal{L}_{\text{NCE}}(\mathbf{f}_{ACN,j}^m, \mathbf{f}_{ACN,j}^{m_i}) =$$
$$\left[-\log\left(\frac{P(\mathbf{f}_{ACN,j}^m | \mathbf{f}_{ACN,j}^{m_i})}{P(\mathbf{f}_{ACN,j}^m | \mathbf{f}_{ACN,j}^{m_i}) + \frac{|\mathcal{N}_j|}{|\mathcal{N}|}} \right) \right.$$
$$\left. + \sum_{k \in \mathcal{N}_j} \log\left(\frac{P(\mathbf{f}_{ACN,k}^m | \mathbf{f}_{ACN,j}^{m_i})}{P(\mathbf{f}_{ACN,k}^{body} | \mathbf{f}_{ACN,j}^{face}) + \frac{|\mathcal{N}_j|}{|\mathcal{N}|}} \right) - 1 \right]$$

(4)

that computes the probability of both features $\mathbf{f}_{ACN,j}^m$ and $\mathbf{f}_{ACN,j}^{m_i}$ representing the same instance u_j compared to other elements in a uniformly sampled negative set $\mathcal{N}_j$. $\mathcal{N}$ is the sample batch. The averaged Focal Loss $\mathcal{L}_{fl}$ [17], specifically effective for an imbalanced dataset like ours, is employed to preserve the category details and a combined loss function $\mathcal{L} = \mathcal{L}_{ACE} + \mathcal{L}_{fl}$ to jointly learn its mode-specific *Central* query networks.

3.5.2. Multimodal Joint Embedding

To address the potential missing modality scenarios in a test environment, we introduce an effective joint latent representation learning model that follows a similar approach as proposed by Radford et al. [18]: the latent descriptor for a positive pair (i.e. a pair of mode-specific representatives describing the same sample using two different modes) are mapped closely, while a negative pair (i.e. a pair of mode-specific representatives describing the two different samples using two different modes) samples will be mapped father from each other in the joint embedding space. Formally, given a minibatch of $\mathcal{N}$ of samples represented using their cross-attended descriptors in a two-mode space (i.e. $\mathcal{N} := \{(\mathbf{f}_i^{m_k}, \mathbf{f}_i^{m_l})\}_{i, m_k \neq m_l}$), we use Jensen-Shannon Divergence, which is a symmetric version of the KL-Divergence to learn the pair-wise invertible linear mappings $S_{m_k \to m_{joint}} : \mathcal{R}^{d_{m_k}} \to \mathcal{R}^{d_{joint}}$,

Table 1. Performance Comparison of different methods using the weighted average F1 measure (W-Avg F1) on the MELD dataset with uni modal (T-Text, A-Audio, and V- Video) and multi-modal representation. Due to the imbalanced class distribution of the dataset, the 'Fear' and 'disgust' classes are represented as the minority classes, the proposed method was also compared against other 5 majority classes ('Neutral', 'Surprise', 'Sadness', 'Joy', and 'Anger') in the dataset and the results are reported in column 'w-avg F1 5 CLS'. More details on emotion-specific comparison are provided in the supplementary material.

Method	Mode	w-Avg F1	w-avg F1 5-CLS
MFN [19]	T + A	0.547	0.5732
ICON [20]	T	0.546	0.5718
	A	0.377	0.3947
	T + A	0.563	0.5897
DialogueRNN [21]	T	0.551	0.5759
	A	0.34	0.3542
	T + A	0.57	0.5971
ConGCN [22]	T	0.574	0.5969
	A	0.422	0.44
	T + A	0.594	0.6175
DialogueCRN [23]	T + A	0.6073	-
EmoCaps [24]	T + A + V	0.6400	-
M2FNet [3]	T + A + V	0.6785	-
Cross-Modal Distribution Matching [25]	T + A	0.571	-
Transformer Based Cross-modality Fusion [26]	T + A + V	0.64	-
Hierarchical Uncertainty [27]	T + A + V	0.59	-
Shape of Emotion [28]	T + A + V	0.63	-
UniMSE [29]	T + A + V	0.66	-
EmotionCLIP [30]	T + A + V	0.3459	-
AM^2-*EmoJE*	T + A	0.6263	0.6845
	T + V	0.6196	0.6782
	A + V	0.5285	0.5825
	T + A (JE)	0.6836	0.7051
	T + V (JE)	0.6897	0.7106
	A + V (JE)	0.6085	0.6572
	No Face	0.6914	0.7944
	No Body	0.7089	0.8117
	T + A + V	**0.7198**	**0.8205**

$\forall m_k \in \mathcal{M}$ to a single joint embedding space $\mathcal{D}_{joint} \subset \mathcal{R}^{d_{joint}}$ and the corresponding loss function is defined as:

$$\mathcal{L}_{m_k \to m_{joint}} = \frac{-1}{|\mathcal{N}|} \Big[\sum_{x \in \mathcal{N}} \frac{1}{2} \log(\text{KL}(S_{m_k \to m_{joint}}(\mathbf{f}_i^{m_k}) || \mathbf{t}_i^{kl})$$
$$+ \log(\text{KL}(S_{m_l \to m_{joint}}(\mathbf{f}_i^{m_l}) || \mathbf{t}_i^{kl}) \Big]$$

(5)

where $x := (\mathbf{f}_i^{m_k}, \mathbf{f}_i^{m_l})$ and $\mathbf{t}_i^{kl} = \frac{S_{m_k \to m_{joint}}(\mathbf{f}_i^{m_k} + \mathbf{f}_i^{m_l})}{2}$. The model jointly optimizes a single loss function $\mathcal{L}_{joint} := \sum_{\substack{m_k, m_l \in \mathcal{M} \\ m_k \neq m_l}} \mathcal{L}_{m_k \to m_{joint}}$ to learn the collection of invertible mapping functions: $\{S_{m_k \to m_{joint}}\}_{m, k \in \mathcal{M}}$.

4. EXPERIMENTS

4.1. Datasets

MELD [1] is a multimodal group conversational dataset that sources clips from the TV Series FRIENDS and consists of the 6 basic emotions. IEMO-CAP [24] is a dual-party conversational dataset acted out by professional actors and features conversations that are far longer on average than those in MELD.

4.2. Results

Tables 1 and 2 report the comparative performance of the proposed method against several state-of-the-art methods [20, 19, 21, 23, 24, 3, 25, 31, 26, 27, 28, 29] in the MELD and IEMOCAP datasets, respectively. While text appears as the most reliable unimodal feature, combining multiple features has

been undoubtedly helpful to **deliver a competitive** 71.98% **weighted F1-score (w-avg F1), which is an improvement of** $\approx 4\%$ **across all classes, as compared to the best-performing baseline M2FNet[3].** MELD is an unbalanced dataset that has 'fear' and 'disgust' identified as the minority classes, with less than 50 samples as representatives. AM^2-*EmoJE* reports 82.05% w-avg F1, which is **a significant** 20% **improvement in the w-avg F1 score compared to the best-performing baseline using the five majority classes.** Compared to the existing works, which offer equal emphasis on all modes, adaptively identifying the more reliable modes to design the relative weight assignments in a query-specific manner has been proven to be effective for delivering a more accurate estimate of a speaker's emotional state. Furthermore, from the performance reported in the sub-row described using the mode 'T+A+V(no face)' (that uses cues from audio, transcript, and visual capturing subjects' body regions only), we observe that the proposed AM^2-*EmoJE* still remains competitive (i.e., reporting $\approx 2\%$ improved w-avg F1 score) against the best-performing baseline. In fact, as we compare the result reported in another sub-row described using mode 'T+A+V(no body)' (that uses cues from audio, transcript, and visual capturing subjects' face regions only), we observe that AM^2-*EmoJE*'s performance remains nearly equivalent as in two different multi-mode data scenarios. Finally, to evaluate the contribution of the proposed multimodal joint learning (Section 3.5.2), in the table we also report the performance of the proposed AM^2-*EmoJE* in multiple missing modality scenarios. **For example, comparing the performance between the sub-row described using mode 'T+A' (that uses only text and audio) and 'T+A(JE)' (that uses only text and audio and uses multimodal joint learning to compensate for the visual modality) we note that the proposed multimodal joint learning module enables around** 5% **(and** 2%**) gain in the w-avg F1 score overall (and five majority) classes.** A similar observation can also be made by comparing the subrows 'A+V' and 'A+V(JE)' (and subrows 'T+V' and 'T+V(JE)') which reports around 8% (and 7%) performance gain in the w-avg F1 score. Augmented with the proposed multimodal joint learning module, AM^2-*EmoJE* attains a state-of-the-art performance in two-mode data scenarios, where the queries are presented with a variety of missing modality scenarios (e.g. 'video' or 'audio').

A similar performance was also observed in Table 2 which reports the comparative analysis using the IEMOCAP dataset. The proposed AM^2-*EmoJE* **reports** 74.91% **w-avg F1, which is around** 4.9% **improvement compared to the best-performing baseline LIGHT-SERNET[32].** Being equipped with the proposed multimodal joint learning module, the model again demonstrates a competitive performance in two-mode data scenarios, where the queries are represented using 'text and audio' or 'text and video'.

Table 3 reports an ablation study. The first two rows in the Table demonstrate the superiority of the weighted multimodal attention (Section 3.2) and weighted fusion techniques (Section 3.3) compared to vanilla cross-attention and static weights (where we use equal weight for all modes). The subrows in the third row report the performance of the proposed AM^2-*EmoJE* where only classification focal loss $\mathcal{L}_{fl}$ is used, but we did not use $\mathcal{L}_{ACE}$ in Eqn (3). **As observed in the table, while the introduction of the Guided NCE (Section 3.5.1) with focal loss delivers around** $1-2\%$ **improvement in w-F1 score, by enabling a query adaptive multimodal fusion scheme (Section 3.3)** AM^2-*EmoJE* **delivers a robust** $(2-4\%)$ **improved) the w-F1 score.**

5. CONCLUSION

We present AM^2-*EMOJE* with our weighted multimodal attention and query adaptive fusion, allowing us to effectively combine the information from various modes to make better decision about the emotional state of subjects in a group conversation. The model is also trained with our proposed *Guided NCE* loss that allows the model to learn representation of the subjects with only their facial features or body language, allowing us to better preserve the privacy of participants and still achieve performance that is very close to the state-of-the-art. Furthermore, we also propose an effective *Multimodal Joint-Embedding* scheme that allows the model to effectively compensate for missing modalities during inference, allowing its performance on a subset of modalities to be close to its performance on the full set of modalities as shown in our comparison with other state-of-the-art models and in the ablation studies in the absence of Joint-Embedding.

Table 2. Performance comparison of difference methods using the weighted average F1 measure (W-Avg F1) on the IEMOCAP dataset with uni (T:=Text, A:=Audio, and V:= Video) and multi-modal Data Representations. 'Feature Concat' in row 13 and row 14 describe the concatenation of multiple uni-mode descriptors to define a multimodal descriptor. More details on emotion-specific comparison are provided in the supplementary material

Method	Mode	w-Avg F1
MFN[19]	T + A	0.3490
ICON[20]	T + A + V	0.6350
DialogueRNN[21]	T + A + V	0.6275
MMGCN[33]	T + A + V	0.6622
DialogueCRN[23]	T + A	0.6620
Hierarchical Uncertainty [27]	T + A + V	0.6598
DAG-ERC+HCL[31]	T	0.6803
M2FNet[3]	T + A + V	0.6986
LIGHT-SERNET[32]	T + A + V	0.7020
AM^2-*EmoJE*	T + A	0.6162
	T + V	0.6343
	A + V	0.5379
	T + A (JE)	0.6919
	T + V (JE)	0.7094
	A + V (JE)	0.6580
	No Face	0.7175
	No Body	0.7286
	T + A + V	0.7491

Table 3. Ablation Study using w-F1 Score

Model	Modalities	MELD	IEMOCAP
No Multimodal Attention	T+A+V	0.6160	0.7527
Equal Fusion Weights	T+A+V	0.6934	0.7165
No Guided NCE	No Body	0.6343	0.6491
	No Face	0.6021	0.6182
	T+A+V	0.7041	0.7286
AM^2-*EmoJE*	T+A+V	0.7198	0.7491

6. REFERENCES

[1] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," 2019.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335, Nov 2008.

[3] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.

[4] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, May 2022, pp. 1610–1618, Association for Computational Linguistics.

[5] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Efficient large-scale multi-modal classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[6] Piao Shi, Min Hu, Fuji Ren, Xuefeng Shi, and Liangfeng Xu, "Learning modality-fused representation based on transformer for emotion analysis," *Journal of Electronic Imaging*, vol. 31, no. 6, pp. 063032, 2022.

[7] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, "Are multimodal transformers robust to missing modality?," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18156–18165.

[8] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee, "Multimodal prompting with missing modalities for visual recognition," *2023 IEEE/CVF Con-

ference on Computer Vision and Pattern Recognition (CVPR), pp. 14943–14952, 2023.

[9] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, pp. 101973, 2024.

[10] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2022, SIGIR '22, p. 1545–1554, Association for Computing Machinery.

[11] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng, "M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. 2023, AAAI'23/IAAI'23/EAAI'23, AAAI Press.

[12] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao, "M3care: Learning with missing modalities in multimodal healthcare data," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2022, KDD '22, p. 2418–2428, Association for Computing Machinery.

[13] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019.

[14] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[15] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer, "Efficient training of audio transformers with patchout," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. 2022, pp. 2753–2757, ISCA.

[16] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi, "Active learning by feature mixing," 2022.

[17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *CoRR*, vol. abs/1708.02002, 2017.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[19] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[20] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct.-Nov. 2018, pp. 2594–2604, Association for Computational Linguistics.

[21] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," 2019.

[22] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. 7 2019, pp. 5415–5421, International Joint Conferences on Artificial Intelligence Organization.

[23] Dou Hu, Lingwei Wei, and Xiaoyong Huai, "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations," *arXiv preprint arXiv:2106.01978*, 2021.

[24] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu, "Emocaps: Emotion capsule based model for conversational emotion recognition," *arXiv preprint arXiv:2203.13504*, 2022.

[25] Jingjun Liang, Ruichen Li, and Qin Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2852–2861.

[26] Baijun Xie, Mariia Sidulova, and Chung Hyuk Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, pp. 4913, 2021.

[27] Feiyu Chen, Jie Shao, Anjie Zhu, Deqiang Ouyang, Xueliang Liu, and Heng Tao Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Transactions on Cybernetics*, 2022.

[28] Harsh Agarwal, Keshav Bansal, Abhinav Joshi, and Ashutosh Modi, "Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts," *arXiv preprint arXiv:2112.01938*, 2021.

[29] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li, "Unimse: Towards unified multimodal sentiment analysis and emotion recognition," *arXiv preprint arXiv:2211.11256*, 2022.

[30] Sitao Zhang, Yimu Pan, and James Z. Wang, "Learning emotion representations from verbal and nonverbal communication," 2023.

[31] Lin Yang, Yi Shen, Yue Mao, and Longjun Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *AAAI Conference on Artificial Intelligence*, 2021.

[32] Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, and Benoit Champagne, "Light-sernet: A lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6912–6916.

[33] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, MM '19, p. 1437–1445, Association for Computing Machinery.