A Unimodal Valence-Arousal Driven Contrastive Learning Framework for Multimodal Multi-Label Emotion Recognition

Wenjie Zheng
School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, China
wjzheng@njust.edu.cn

Jianfei Yu*
School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, China
jfyu@njust.edu.cn

Rui Xia*
School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, China
rxia@njust.edu.cn

Abstract

Multimodal Multi-Label Emotion Recognition (MMER) aims to identify one or more emotion categories expressed by an utterance of a speaker. Despite obtaining promising results, previous studies on MMER represent each emotion category using a one-hot vector and ignore the intrinsic relations between emotions. Moreover, existing works mainly learn the unimodal representation based on the multimodal supervision signal of a single sample, failing to explicitly capture the unique emotional state of each modality as well as its emotional correlation between samples. To overcome these issues, we propose a **Uni**modal **V**alence-**A**rousal driven contrastive learning framework (UniVA) for the MMER task. Specifically, we adopt the valence-arousal (VA) space to represent each emotion category and regard the emotion correlation in the VA space as priors to learn the emotion category representation. Moreover, we employ pre-trained unimodal VA models to obtain the VA scores for each modality of the training samples, and then leverage the VA scores to construct positive and negative samples, followed by applying supervised contrastive learning to learn the VA-aware unimodal representations for multi-label emotion prediction. Experimental results on two benchmark datasets MOSEI and M^3ED show that the proposed UniVA framework consistently outperforms a number of existing methods for the MMER task. The source code is publicly released at https://github.com/NUSTM/UniVA.

CCS Concepts

• **Information systems** → **Sentiment analysis**; • **Computing methodologies** → **Natural language processing**.

Keywords

Multimodal Emotion Recognition; Multimodal Multi-Label Learning; Contrastive Learning

*Corresponding author.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681638

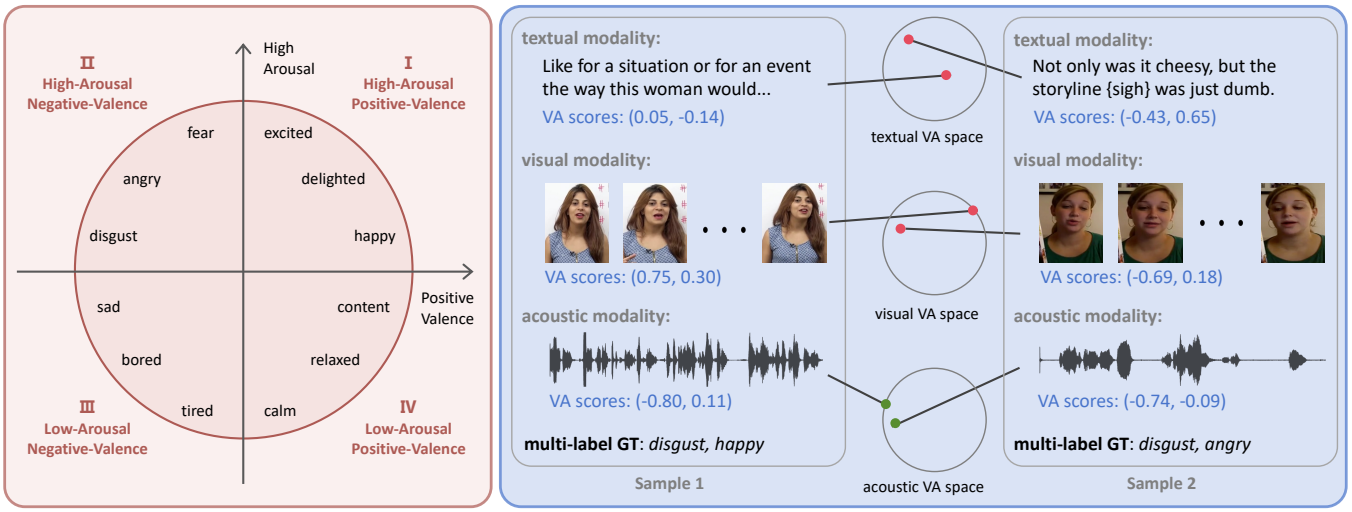
ACM Reference Format:
Wenjie Zheng, Jianfei Yu, and Rui Xia. 2024. A Unimodal Valence-Arousal Driven Contrastive Learning Framework for Multimodal Multi-Label Emotion Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681638

1 Introduction

Multimodal Emotion Recognition has recently attracted considerable attention [14], as emotions play a great impact on human cognition, decision-making, and social interactions. Given that an utterance of a speaker in conversations or videos may naturally express more than one emotion category, recent studies attempt to explore the Multimodal Multi-label Emotion Recognition (MMER) task, which aims to integrate multimodal information sources, i.e., text, vision, and audio, to identify one or more emotion categories expressed by an utterance of a speaker such as *happy* and *angry* [15, 24].

Existing studies typically model the MMER task as a multi-label classification problem. One line of work focuses on designing different inter-modal interaction mechanism to obtain the multimodal representation and capturing its dependency on each emotion category [15, 64, 65]. Another line of work focuses on modeling the dependency between the emotion categories by proposing a label-aware Transformer decoder [24] or different multi-label loss functions [3, 15].

Despite obtaining promising results on several benchmark datasets for the MMER task, most existing studies still suffer from several limitation. First, most existing MMER studies [24, 63] represent each emotion category using a one-hot vector and regard it as an independent label, ignoring the intrinsic relationship between different emotion categories. For example, *happiness* and *surprise* are often encoded as distinct positive emotions, without considering their shared characteristic of conveying high emotional intensity. Similarly, *sadness* and *boredom* are both encoded as distinct negative labels, ignoring their commonality in terms of lower emotional intensity. Second, some studies [18, 61, 65] have recognized the importance of learning a modality-specific representation for each modality, e.g., Yu et al. [61] utilize multimodal annotations to generate unimodal labels. However, they primarily learn the representation of each modality based on the multimodal supervision signals, failing to explicitly capture the unique emotional state of each modality. For example, in Sample 1 of Fig. 1, although the



(a) Valence-Arousal space (b) Two examples of MMER

Figure 1: For VA space, based on the positivity or negativity of valence and the high or low levels of arousal, it is divided into four quadrants, each containing several discrete emotion categories.

multi-label ground truth is (*disgust*, *happy*) and the visual modality clearly displays *happy*, the emotion displayed by the textual modality tends towards *neutral*. Such unimodal emotional state, used to mitigate the polarity of emotions expressed by other modality and to prevent prediction biases caused by overreliance on the polarized emotional modality, is difficult to obtain solely relying on multimodal supervision signals. The work [60] attempted to manually annotate each unimodal label for the single-label emotion recognition task. Nevertheless, it leads to a high cost. Lastly, existing methods mainly focus on learning the multimodal representation with the supervision signal of a single sample, ignoring the emotion correlations between different samples. For instance, as shown in the acoustic modality of Fig. 1, if two samples have similar emotional states in one modality, their representations in that modality tend to be similar.

To address the above limitations, we propose a **Uni**modal **V**alence-**A**rousal driven contrastive learning framework named UniVA for the MMER task. Specifically, to capture the intrinsic relationship between emotion categories, we adopt the widely-used valence-arousal (VA) space [5] to represent each emotion category with a dimensional *valence* score and a dimensional *arousal* score. As illustrated in Figure 1 (a), the *valence* measures the positivity or negativity of an emotion and the *arousal* indicates its intensity, which well map emotions in a manner that reflects their inherent similarities or differences. Thus, we derive the correlation between emotion categories from the VA space and use it as priors to learn the emotion category representation. Secondly, to explicitly model the emotional state of each modality, we propose to obtain the *valence* and *arousal* scores for each modality based on the unimodal models that are pre-trained on existing VA datasets. By utilizing VA scores, we can gain a detailed understanding of the emotional dynamics among different modalities and how each modality contributes to and influences the multimodal prediction. Moreover, to consider the emotion correlations among different samples, we first measure the similarity between each modality of a pair of training samples based on their unimodal VA scores, and then leverage the similarity score to construct positive and negative sample pairs for each sample. With the positive and negative sample pairs, we apply a supervised contrastive learning model to obtain the VA-aware unimodal representations, and integrate them as the multimodal representation for multi-label emotion prediction.

The main contributions in this work can be summarized as follows:

- We propose to represent each emotion category with the valence-arousal (VA) space to capture the correlation between emotion categories and use it as priors to learn the emotion category representation for the MMER task.
- We design a unimodal VA-driven contrastive learning algorithm, which first obtains the VA scores for each modality based on pre-trained models, and then utilize these VA scores to construct positive and negative samples for supervised contrastive learning.
- Extensive evaluation on two benchmark MMER datasets MOSEI and M^{3}ED demonstrate the superiority of the proposed framework UniVA over many previous multimodal methods and the effectiveness of each component in UniVA on different multi-label evaluation metrics.

2 Related Work

2.1 Emotion Recognition

Single-Label Emotion Recognition (SLER) is an important task in the field of affective computing. According to input sources, SLER is divided into textual SLER and multimodal SLER. For textual SLER, modeling contextual dependencies has become a widely discussed topic for emotion recognition in conversations [38, 48]. Some works [36, 47] also attempt to model speaker dependencies. Moreover, researchers [16, 30] are interested in improving performance by introducing commonsense and analyzing the speaker's

mental states. With the emergence of large language models (LLMs) like ChatGPT [54, 67], there has been a series of works that combine these LLMs [29, 34]. Recently, the development of multimedia has drawn attention to multimodal SLER [45]. Some researchers focus on the importance of different modalities [25] and challenge of multimodal fusion [21, 42]. Also, some works focus on the field of conversation [17, 22], and some researchers are focusing on proposing robust approaches [20, 31].

Multi-Label Emotion Recognition (MLER) is a task of identifying one or more emotions in a given text or video. Existing studies can be divided into Textual MLER and Multimodal MLER. Firstly, for Textual MLER [11, 71], considering intrinsic relations between emotions, Wang and Zong [58] and Huang et al. [23] model emotional dependencies within text representations. Meanwhile, such as Fei et al. [12] and Ma et al. [37], focus on distinguishing similar labels and learning distinct semantic representations for different labels. Furthermore, Fei et al. [13] consider the prior emotion distribution in sentences and capture the context information relevant to those emotions. Recently, some researches, such as [24, 63–65], have begun to delve into Multimodal MLER. Akhtar et al. [1] design a multi-task learning approach to enhance performance of model. Anand et al. [3] propose multimodal distillation loss to improve the generalization ability. Srivastava et al. [49] design multimodal method to understand emotions and mental states of characters in movie scenes.

2.2 Valence-Arousal Application

In the field of affective computing, application of multi-dimensional valence and arousal [19, 72] is increasingly widespread due to its ability to provide a more detailed understanding of emotions, compared to discrete emotion categories. To further explore the correlation between discrete and dimensional emotions, several studies have introduced datasets for multi-task learning, such as IEMOCAP [7] and MER2023 [32]. With these datasets, many multi-task learning methods have been proposed by [8, 43]. Considering the broad application prospects of continuous emotion prediction in real scenarios, various workshops and competitions have been introduced, such as AVEC [46, 55], MuSe [2], and ABAW [27]. Moreover, some works have employed the NRC-VAD lexicon [40] as an external knowledge base for the emotion recognition task [59, 69] and the empathetic response generation task [9, 70].

3 Methodology

In this section, we first introduce the task definition and the overview of our UniVA framework. We then describe the details of each module in UniVA.

3.1 Task Definition and Framework Overview

Given a MMER corpus $\mathbb{D} = \left\{(u^i, y_i)\right\}_{i=1}^{N}$, the input of each sample $u^i = \left\{u_t^i, u_v^i, u_a^i\right\}$ is an utterance that contains information from three modalities, i.e., text, vision, and audio, denoted by $\{t, v, a\}$. The output $y_i = \left\{y_i^1, y_i^2, ..., y_i^C\right\}$ is a pre-defined label sequence with C emotions, where $y_i^j \in \{0, 1\}$ indicates whether or not u^i contains the j-th emotion. The goal of MMER task is to learn a mapping function $\mathcal{F} = (u_t^i, u_v^i, u_a^i) \rightarrow y_i$ to predict the occurrence of each emotion category.

Figure 2 shows the overview of UniVA that contains three key modules, i.e., VA Scores Acquisition, VA-Driven Contrastive Learning-based Unimodal Representation, and Multi-Label Prediction with VA-Driven Emotion Correlation Priors. Specifically, we adopt the widely-used VA space [5], and use either a NRC-VAD lexicon [40] or pre-trained VA models to obtain the VA scores for each emotion category and each modality of the training samples. The second module then leverages the VA scores to construct positive and negative sample pairs for each sample, which are then used to train a supervised constrastive learning model to obtain the VA-aware unimodal representations. Lastly, the third module integrates the unimodal representations and incorporates the correlation prior between emotion categories in the VA space as a regularization term for multi-label emotion prediction.

3.2 VA Scores Acquisition

Given an utterance u^i and its multi-label annotation y_i, we obtain the VA scores for y_i and three modalities $\left\{u_t^i, u_v^i, u_a^i\right\}$ as follows:

Label. Given an emotion y_i^j of y_i, we directly obtain the valence and arousal scores $(\mathcal{V}_e^j, \mathcal{A}_e^j)$ from the NRC-VAD lexicon [40], which provides reliable human ratings of valence, arousal, and dominance for 20,000 English terms.

Text. We fine-tune a RoBERTa-base model [33] on the EmoBank dataset [6] and feed the textual input u_t^i into the model for inference. We then obtain a valence score $\mathcal{V}_t^i \in [-1, 1]$ and an arousal score $\mathcal{A}_t^i \in [-1, 1]$.

Vision. For the visual input u_v^i, we first extract its facial sequence $\boldsymbol{s^i}$ and feed it into the EmoFAN model [52], which has been trained on the AffectNet dataset [41]. We then obtain a valence score and an arousal score for each face, and average the valence and arousal scores across the facial sequence to derive the overall valence score $\mathcal{V}_v^i \in [-1, 1]$ and arousal score $\mathcal{A}_v^i \in [-1, 1]$.

Audio. For the acoustic input u_a^i, we feed it into the Wav2Vec2-Large-Robust model [57], which was fine-tuned on the MSP-Podcast dataset [35] and has been shown to exhibit excellent generalization and robustness, to obtain a valence score $\mathcal{V}_a^i \in [-1, 1]$ and an arousal score $\mathcal{A}_a^i \in [-1, 1]$.

3.3 VA-Driven Contrastive Learning-based Unimodal Representation

In this subsection, we introduce the details of learning the unimodal representation based on VA-Driven Contrastive Learning.

3.3.1 Unimodal Feature Extraction. For an utterance $u^i = \left\{u_t^i, u_v^i, u_a^i\right\}$, we employ existing feature extraction methods to obtain the textual, visual, and acoustic features, i.e., $\mathbf{X}_t \in \mathbb{R}^{l_t \times k_t}$, $\mathbf{X}_v \in \mathbb{R}^{l_v \times k_v}$, and $\mathbf{X}_a \in \mathbb{R}^{l_a \times k_a}$. Here $l_{m \in \{t,v,a\}}$ denotes the sequence length of each modality, and $k_{m \in \{t,v,a\}}$ is the feature dimension.

Specifically, for the textual input u_t^i, we utilize either Glove [44] or RoBERTa [33] to obtain the word representation. For Glove, we directly input u_t^i to obtain the text representation $\mathbf{X}_t$. For RoBERTa, we first concatenate the text from all clips in the current video by inserting special tokens $\langle/s\rangle$, and then feed them into the model to obtain $\mathbf{X}_t$.

For the audio input u_a^i sampled at 16kHz, Wav2Vec2.0 model [10] is utilized to extract low-level acoustic features $\mathbf{X}_a$.

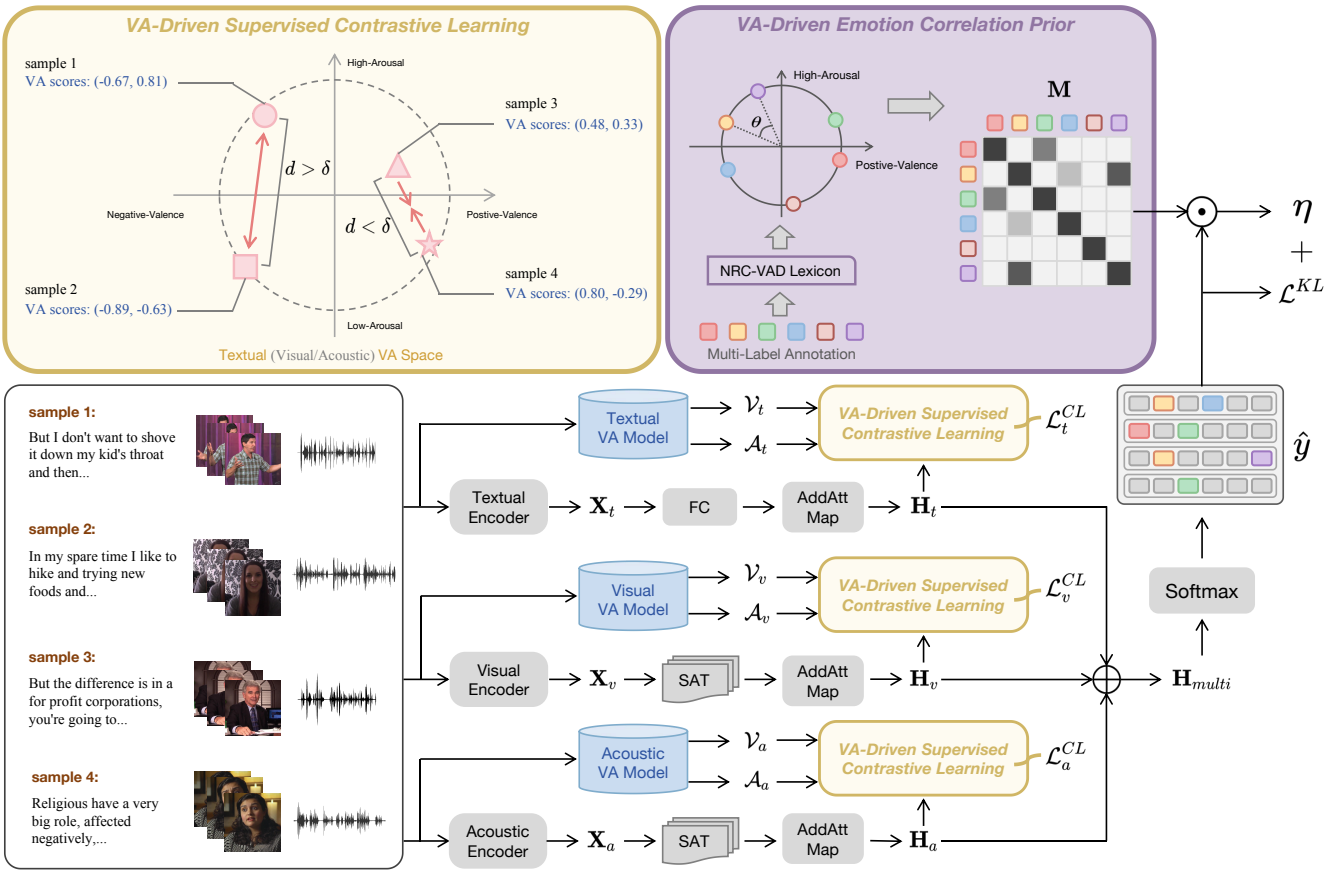


Figure 2: The overview of our proposed Unimodal Valence-Arousal driven contrastive learning framework (UniVA).

For the video clip u_v^i, we first employ the method [68] to extract facial sequence $\boldsymbol{s^i} = \left\{s_1^i, s_2^i, ..., s_q^i\right\}$, where q denotes the total number of faces. These facial images are then fed to the Inception-ResNetv1 model [50] to obtain frame-level visual features $\mathbf{X}_v$.

Intra-Modal Interaction. For the textual modality, we employ a fully connected (FC) layer and additive attention (AddAtt) mapping [4] to obtain the utterance-level representation:

$$\mathbf{H}_t = \text{AddAtt}(\text{FC}(\mathbf{X}_t)), \tag{1}$$

where $\mathbf{H}_t \in \mathbb{R}^{d_t}$ and d_t is the hidden dimension.

For visual and acoustic modalities, we respectively feed $\mathbf{X}_v$ and $\mathbf{X}_a$ into two separate Self-Attention (SAT) layers, followed by the additive attention mapping to obtain the utterance-level visual and acoustic representations as follows:

$$\mathbf{H}_{m\in\{v,a\}} = \text{AddAtt}(\text{SAT}(\mathbf{X}_{m\in\{v,a\}})), \tag{2}$$

where $\mathbf{H}_v \in \mathbb{R}^{d_v}$ and $\mathbf{H}_a \in \mathbb{R}^{d_a}$.

3.3.2 VA-Driven Contrastive Learning. Inspired by the supervised contrastive learning (SupCon) introduced by Khosla et al. [26], we utilize the VA scores of each modality as supervision signals to consider the relationship between samples to enhance the unimodal representation.

Positive and Negative Sample Construction. For an anchor sample $\boldsymbol{x}^i$, the key question in supervised contrastive learning is how to obtain samples semantically similar to (or different from) $\boldsymbol{x}_i$, which are called *positive* samples $\boldsymbol{x}_i^+$ (or *negative* samples $\boldsymbol{x}_i^-$). In previous studies, since SupCon is applied in the single-label classification task, we can obtain positive samples $\boldsymbol{x}_i^+$ and negative samples $\boldsymbol{x}_i^-$ based on the labels of samples. However, since there are many co-occurred emotions in the MMER task, it is hard to construct $\boldsymbol{x}_i^+$ and $\boldsymbol{x}_i^-$ based on the emotion labels.

The VA space provides a rich, continuous spectrum of emotional states, allowing for a more precise and meaningful categorization of emotional similarity and difference. Therefore, we propose to utilize the VA scores of each modality to construct the positive and negative samples. In this way, positive samples $\boldsymbol{x}_i^+$ are not merely those sharing the same categorical label with the anchor, but rather those whose VA scores indicate a close emotional proximity. Conversely, negative samples $\boldsymbol{x}_i^-$ are identified through significant divergences in their VA scores from the anchor, reflecting a fundamental emotional disparity. This method acknowledges the multidimensional nature of emotions, recognizing that two samples could share a label (e.g., *happy*) while embodying different intensities or nuances of that emotion.

Specifically, for any modality $m \in \{t, v, a\}$, assuming the batch size is B, we are given two samples u_m^i and u_m^j, where $i, j \in B$, and their VA scores are $(\mathcal{V}_m^i, \mathcal{A}_m^i)$ and $(\mathcal{V}_m^j, \mathcal{A}_m^j)$, respectively. We first measure their similarity based on their Euclidean distance in

the VA space below:

$$d(u_m^i, u_m^j) = \sqrt{(\mathcal{V}_m^i - \mathcal{V}_m^j)^2 + (\mathcal{A}_m^i - \mathcal{A}_m^j)^2}. \tag{3}$$

Based on the similarity score d, we then determine whether the two samples form a positive or negative pair with a predefined threshold δ. If $d < \delta$, u_m^i and u_m^j are considered as a positive pair; otherwise, they are deemed as a negative pair.

To prevent the scenario where a batch consists entirely of negative pairs, we duplicate $\mathbf{H}_{m\in\{t,v,a\}}$ and obtain multi-view unimodal representations $\tilde{\mathbf{H}}_m = [\mathbf{H}_m, \mathbf{H}_m]$. Finally, for each anchor sample $\boldsymbol{x}_i \in \boldsymbol{X} \equiv \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_B\}$, the loss of the VA-driven contrastive learning is defined as follows:

$$\mathcal{L}_m^{CL} = \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \frac{-1}{|P(\boldsymbol{x}_i)|} \sum_{\boldsymbol{x}_p \in P(\boldsymbol{x}_i)} \text{sim}(\boldsymbol{x}_p, \boldsymbol{x}_i), \tag{4}$$

$$\text{sim}(\boldsymbol{x}_p, \boldsymbol{x}_i) = \log \frac{\exp((\tilde{\mathbf{H}}_m^i \cdot \tilde{\mathbf{H}}_m^p)/\tau)}{\sum_{\boldsymbol{x}_a \in A(\boldsymbol{x}_i)} \exp((\tilde{\mathbf{H}}_m^i \cdot \tilde{\mathbf{H}}_m^a)/\tau)}, \tag{5}$$

where $P(\boldsymbol{x}_i) = \{\boldsymbol{x}_j \in A(\boldsymbol{x}_i) \mid d(u^i, u^j) < \delta, j \neq i\}$ represents the set of all positive samples paired with anchor $\boldsymbol{x}_i$, $A(\boldsymbol{x}_i) \equiv \boldsymbol{X} \setminus \{\boldsymbol{x}_i\}$, and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter.

3.4 Multi-Label Prediction with VA-Driven Emotion Correlation Priors

After obtaining the VA-aware unimodal representations, we concatenate them as the multimodal representation $\mathbf{H}_{multi}$, and then feed $\mathbf{H}_{multi}$ into a softmax layer to obtain the emotion distribution $\hat{y}$ for multi-label emotion prediction:

$$\mathbf{H}_{multi} = \text{Concat}(\mathbf{H}_t, \mathbf{H}_v, \mathbf{H}_a), \tag{6}$$

$$\hat{y} = \text{Softmax}(\mathbf{W}^{\mathrm{T}}\mathbf{H}_{multi} + b), \tag{7}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters.

VA-Driven Emotion Correlation Priors. To further capture the intrinsic relationship between emotion categories, we introduce a VA-driven emotion correlation prior η for $\hat{y}$. Specifically, assuming we have C emotions, we first calculate their similarity in the VA space to obtain the emotion similarity matrix $\mathbf{M} \in \mathbb{R}^{C\times C}$ as follows:

$$\mathbf{M}_{jl} = \frac{\mathcal{V}_e^j \cdot \mathcal{V}_e^l + \mathcal{A}_e^j \cdot \mathcal{A}_e^l}{\sqrt{(\mathcal{V}_e^j)^2 + (\mathcal{A}_e^j)^2} \cdot \sqrt{(\mathcal{V}_e^l)^2 + (\mathcal{A}_e^l)^2}} \tag{8}$$

where $\left(\mathcal{V}_e^j, \mathcal{A}_e^j\right)$ and $\left(\mathcal{V}_e^l, \mathcal{A}_e^l\right)$ respectively denote the VA scores of the j-th emotion and the l-th emotion. We then incorporate the emotion correlation prior into our model with the following loss η:

$$\eta = \frac{1}{N} \sum_{i=1}^{N} \sum_{j,l} \mathbf{M}_{j,l} \left\|\hat{y}_{i,j} - \hat{y}_{i,l}\right\|_2^2 \tag{9}$$

where N represents the number of total samples in the training set. During the training process, we aim to minimize this η with the goal of making label predictions on similar emotion positions more similar, and those on dissimilar emotion positions more distinct.

Table 1: The statistics of two benchmark datasets.

Dataset	Split			Multi-Label	
	Train	Valid	Test	One	Two & more
MOSEI	16,326	1,871	4,659	14,517	8,339
M^{3}ED	17,425	2,821	4,201	21,791	2,656

3.5 Model Training

For the main MMER task, we use KL divergence [28] as the multi-label loss function:

$$\mathcal{L}^{KL}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{i=1}^{C} \mathbf{y}^i \log(\frac{\mathbf{y}^i}{\hat{\mathbf{y}}^i}) \tag{10}$$

where $\hat{\mathbf{y}}$ denotes the model prediction, $\mathbf{y}$ denotes the ground truth distribution. The full objective function of our UniVA framework is a combination of the contrastive learning loss, the main task loss, and the emotion correlation prior as follows:

$$\mathcal{L} = \lambda \cdot \textstyle\sum_{m\in\{t,v,a\}} \mathcal{L}_m^{CL} + (1-\lambda) \cdot \mathcal{L}^{KL} + \eta, \tag{11}$$

where λ is a trade-off parameter.

4 Experiments

4.1 Experimental Setup

Datasets. To validate the effectiveness of our framework *UniVA*, we conduct experiments on two benchmark datasets: MOSEI [62] and M^{3}ED [66]. MOSEI has 22,856 utterance-level video clips acquired from YouTube. Each video clip is annotated with either one or more of Ekman's six basic emotions (i.e., *happy*, *sad*, *anger*, *surprise*, *disgust*, and *fear*) or the *neutral* emotion. M^{3}ED contains 24,447 utterances collected from 56 Chinese TV series. It is annotated with six basic emotion categories and an additional *neutral*. Table 1 shows the statistics of the samples with multiple labels of both datasets. Moreover, we introduce the three datasets used during the VA scores acquisition phase. EmoBank is a corpus focused on social media, consisting of 10,000 English sentences, each annotated with valence and arousal. AffectNet is a large facial imagery dataset containing over a million images, each face annotated with valence and arousal scores. MSP-Podcast is a speech emotional dataset containing over 150,000 speech segments from podcast recordings, with each segment annotated for valence and arousal scores.

Implementation Details. For our *UniVA* framework, we employ either Glove-300d or RoBERTa-base as the textual encoder. For M^{3}ED, we use RoBERTa-base in Chinese[1]. The visual encoder InceptionResNet was fine-tuned on the CASIA-WebFace dataset. For the acoustic modality, the acoustic encoder Wav2vec-English[2] used for MOSEI was fine-tuned on the Common Voice 6.1 dataset. Similarly, Wav2vec-Chinese[3] employed for M^{3}ED was fine-tuned using the Common Voice 6.1, CSS10, and ST-CMDS datasets. Given an utterance in M^{3}ED, since our textual VA model is trained on English corpus, we translate the text into English using DeepL API[4], and then feed it into the VA model to obtain the textual VA scores.

[1] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[2] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english
[3] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn
[4] https://www.deepl.com/pro-api?cta=header-pro-api

Table 2: Comparison results of different methods on the MOSEI and M^{3}ED datasets. The baselines tagged with ♠ utilize Glove as textual encoders, while those tagged with ♣ employ RoBERTa as textual encoders. Moreover, the baseline tagged with ⋆ only uses textual and visual modalities, while other models use three modalities. The best results are marked in bold, while the second best results are underlined.

Methods	MOSEI				M^{3}ED			
	Acc (↑)	*HL* (↓)	*miF1* (↑)	*maF1* (↑)	*Acc* (↑)	*HL* (↓)	*miF1* (↑)	*maF1* (↑)
MuIT♠ (Tsai et al. [53])	44.5	0.190	53.1	34.4	-	-	-	-
M3ER♠ (Mittal et al. [39])	40.9	0.195	51.9	34.9	-	-	-	-
HHMPN♠ (Zhang et al. [64])	45.9	0.189	55.6	<u>43.0</u>	-	-	-	-
TAILOR♠ (Zhang et al. [65])	43.7	0.206	49.7	37.1	-	-	-	-
RobMMR♠ (Ge et al. [15])	48.4	<u>0.185</u>	56.9	41.7	-	-	-	-
MDI♣ (Zhao et al. [66])	49.9	0.186	50.2	10.9	47.6	0.159	<u>51.9</u>	33.6
FacialMMT♣ (Zheng et al. [68])	<u>50.1</u>	0.190	<u>59.1</u>	40.8	<u>48.7</u>	<u>0.154</u>	51.7	<u>37.9</u>
Gemini (zero-shot)⋆ (Team et al. [51])	11.2	0.268	23.9	20.6	18.6	0.198	24.1	19.1
UniVA-Glove	49.2	0.205	57.2	37.2	46.4	0.159	49.1	24.2
UniVA-RoBERTa	**51.3**	**0.182**	**60.5**	**44.4**	**50.6**	**0.149**	**53.4**	**40.2**

The batch size for MOSEI and M^{3}ED is set to 12 and 22, respectively. The learning rate and the hidden size in each modality are set to $5e-5$ and 768. The threshold δ for Euclidean distance of contrastive learning is set to 0.1. During inference, we set an inference threshold ζ to 0.18 so that the emotion with scores higher than ζ is predicted as 1. Following previous works [64], we adopt multi-label Accuracy (*Acc*), Hamming Loss (*HL*), Micro-F1 (*miF1*), and Macro-F1 (*maF1*) scores as our evaluation metrics. We optimize parameters with the AdamW optimizer and train our model on 4 *NVIDIA RTX3090* GPUs, each epoch took roughly 4 minutes to run over the dataset, which contain around 17,000 samples.

4.2 Comparison Methods

We compare the proposed framework *UniVA* with the following systems: *MulT* [53] is a multimodal fusion algorithm that does not require modality-aligned inputs and captures inter-modal interactions with Cross-Modal Transformer. *M3ER* [39] uses canonical correlational analysis and multiplicative fusion for multimodal emotion recognition. *HHMPN* [64] models the feature-to-label, modality-to-label, and label-to-label dependencies via heterogeneous graph message passing. *TAILOR* [65] enhances the multimodal diversity with adversarial learning to obtain the shared and private representations of each modality. *RobMMR* [15] introduces two adversarial training strategies, temporal masking and parameter perturbation, to learn a more robust multimodal representation. *MDI* [66] considers emotional dependency of context in dialogues and proposes a dialogue-aware interaction framework. *FacialMMT* [68] improves the importance of visual modality by extracting the facial sequence of the real speaker in conversations. *Gemini*[5] [51] is a large multimodal model that exhibits remarkable capabilities in multimodal understanding.

Note that since *MDI* and *FacialMMT* is designed for the single-label emotion recognition task, we replace the Cross-Entropy loss used in these methods with the same KL loss as used in our approach. For *Gemini*, we first extract five video frames from each video clip, and then feed these into *Gemini-Vision* to obtain video captions with emotion, which are then concatenated with the textual and spoken content and fed into *Gemini* to generate one or more emotions from the pre-defined emotion list.

[5]In this work, the Pro version is used.

4.3 Main Results

In Table 2, we report the results of *UniVA* and all comparison methods on the two datasets. First, we can find that the performance of multimodal fusion methods such as *MuIT* and *M3ER* is relatively poor due to their insufficient consideration of the dependencies among emotions. *HHMPN* and *TAILOR* achieve better results, because of modeling the both modality-emotion and emotion-emotion dependencies in their models. Moreover, *RobMMR* which focuses on the model robustness attains on the best performance on the *HL* metric, while *MDI* and *FacialMMT* achieve significant improvements on metrics like *Acc* and *miF1*. In addition, the performance of *Gemini* is rather limited, revealing that existing large multimodal models may not be suitable for the multi-label emotion recognition task due to the complexity of the task. Lastly, it is clear that *UniVA-RoBERTa* consistently achieves the best performance across all four metrics on both datasets, which demonstrates the effectiveness of our proposed model. Additionally, we find that using Glove instead of RoBERTa as the textual encoder leads to a decrease in performance. When compared with baselines that utilize Glove for text encoding, although slightly inferior to *RobMMR* and *HHMPN* on the MOSEI dataset in the *HL* and *maF1* metrics respectively, the proposed *UniVA-Glove* still achieves certain advantages overall.

4.4 Ablation Study

Effect of Each Component. Firstly, we conduct ablation studies on two main components proposed in *UniVA*. As shown in Table 3, removing the VA-driven contrastive learning (*VA-CL*) results in an average reduction of 1.22 percentage points, especially in *Acc* and *miF1*, with an average decrease of 1.50 percentage points and 1.55 percentage points, respectively. It indicates that by capturing the unique emotional state of each modality and the emotional correlations between samples, *VA-CL* enhances the emotion recognition

Table 3: Ablation study of our *UniVA* framework. *VA-CL* denotes VA-Driven contrastive learning, and *VA-ECP* denote VA-Driven emotion correlation prior.

Methods	MOSEI				M^{3}ED			
	Acc (↑)	*HL* (↓)	*miF1* (↑)	*maF1* (↑)	*Acc* (↑)	*HL* (↓)	*miF1* (↑)	*maF1* (↑)
UniVA	**51.3**	**0.182**	**60.5**	**44.4**	**50.6**	**0.149**	**53.4**	**40.2**
- w/o VA-CL	50.0	0.189	59.3	43.2	48.9	0.157	51.5	39.2
- w/o VA-ECP	51.0	0.186	59.8	42.8	49.1	0.154	51.7	38.2
- w/o VA-CL, VA-ECP	49.7	0.189	58.2	39.0	48.0	0.160	51.4	37.6

Table 4: Ablation study of *UniVA* on different modalities for MOSEI and M^{3}ED.

Methods	MOSEI		M^{3}ED	
	Acc	*miF1*	*Acc*	*miF1*
UniVA	**51.3**	**60.5**	**50.6**	**53.4**
- w/o Vision	51.1	59.9	49.2	52.0
- w/o Audio	49.7	58.1	48.4	51.4
- w/o Vision, Audio	50.6	59.4	48.0	51.6
- w/o Text, Vision	46.7	54.5	38.8	41.3
- w/o Text, Audio	42.3	48.2	40.8	40.7

Table 5: Comparison results of UniVA with previous methods across different modalities for MOSEI and M^{3}ED.

Methods	Modality	MOSEI		M^{3}ED	
		Acc	*miF1*	*Acc*	*miF1*
UniVA (ours)	text	**50.6**	**59.4**	**48.0**	**51.6**
RobMMR		46.2	53.0	43.7	44.1
FacialMMT		48.9	58.3	46.7	49.8
UniVA (ours)	audio	**46.7**	**54.5**	**38.8**	**41.3**
RobMMR		40.2	47.3	33.2	33.7
FacialMMT		45.3	53.7	37.1	39.8
UniVA (ours)	vision	**42.3**	48.2	**40.8**	**40.7**
RobMMR		41.9	**48.5**	40.6	40.5
FacialMMT		40.6	47.7	39.4	39.5

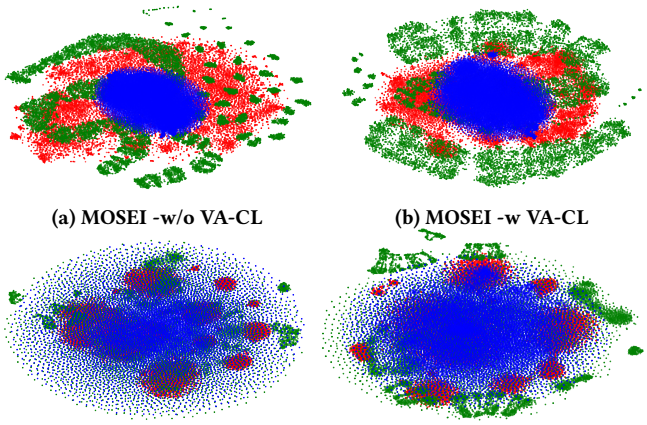
capability. Furthermore, removing the VA-driven Emotion Correlation Prior (*VA-ECP*) leads to a decline in performance across all metrics, particularly on the *maF1*, where there is a decrease of 1.60 percentage points on the MOSEI and 2.00 percentage points on the M^{3}ED. It suggests that *VA-ECP* strengthens the intrinsic relations between different emotion categories.

Effect of Each Modality. We report the results of *UniVA* of removing each modality in Table 4. It is evident that removing one or two modalities consistently leads to the performance drop, indicating that each modality is indispensable for emotion prediction. Among the three modalities, we find that the textual modality is much more important than the other two on both datasets. Moreover, we report the results of UniVA and two strong baselines for the text, audio, and vision modalities in Table 5.

Effect of Different Contrastive Learning. We compared our proposed *VA-CL* with Supervised Contrastive Learning (*Sup-CL*) and Self-supervised Contrastive Learning (*Self-CL*). In *Sup-CL*, we determine positive and negative sample pairs based on whether samples have exactly the same labels: samples with the same labels constitute positive pairs, otherwise they form negative pairs. As shown in Table 6, the results indicate that *VA-CL* outperforms the other methods and demonstrate the effectiveness of *VA-CL*. Moreover, we observed that *UniVA_Sup-CL* performs better than *UniVA_Self-CL*, showing that utilizing label information to differentiate between positive and negative sample pairs is beneficial.

Table 6: Ablation study of *UniVA* with different contrastive learning algorithm. The "-r" indicates that the proposed *VA-CL* is replaced with other algorithms.

Methods	MOSEI		M^{3}ED	
	Acc	*miF1*	*Acc*	*miF1*
UniVA (VA-CL)	**51.3**	**60.5**	**50.6**	**53.4**
-r Sup-CL	50.3	59.0	48.8	52.7
-r Self-CL	49.6	58.2	46.2	50.1



(a) MOSEI -w/o VA-CL (b) MOSEI -w VA-CL

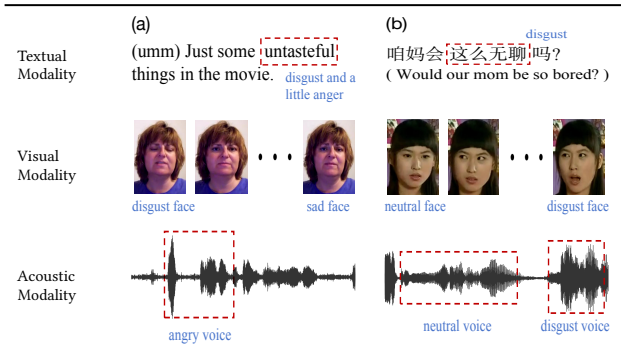
(c) M^{3}ED -w/o VA-CL (d) M^{3}ED -w VA-CL

Figure 3: 2D visualization of each modality on the training set for MOSEI and M^{3}ED: (a)(c) and (b)(d) respectively display unimodal representations without/with *VA-CL*. Red, green, and blue circles denote text, vision, and audio modalities, respectively.

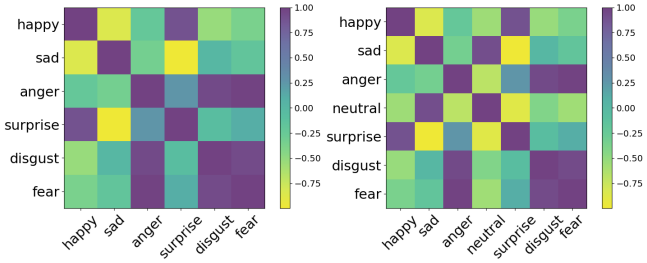
4.5 In-Depth Analysis

Visualization of VA-Aware Unimodal Representations. To demonstrate the effectiveness of the *VA-CL* algorithm, we visualize

Table 7: Prediction comparison on two samples from the test sets of MOSEI and M^{3}ED.



	(a)	(b)
GT	(*disgust*, *angry*, *sad*)	(*neutral*, *disgust*)
TAILOR	(disgust, angry) ✗	(disgust) ✗
FacialMMT	(disgust, sad) ✗	(neutral, disgust) ✓
UniVA (ours)	$(VA)_t$ scores: (-0.43, 0.08) $(VA)_v$ scores: (-0.36, -0.02) $(VA)_a$ scores: (-0.67, 0.51) (disgust, angry, sad) ✓	$(VA)_t$ scores: (-0.17, -0.33) $(VA)_v$ scores: (-0.09, -0.28) $(VA)_a$ scores: (-0.23, -0.31) (neutral, disgust) ✓



(a) MOSEI (b) M^{3}ED

Figure 4: The heatmap of VA-driven emotion correlation matrix on the two benchmark datasets.

the unimodal representations on the training sets of MOSEI and M^{3}ED by t-SNE [56]. As shown in Figure 3, we can observe that with the help of *VA-CL*, the representations of each modality become more distinguishable, and samples within the same modality are more clustered. Specifically, compared to subfigure (a), in subfigure (b) with the aid of *VA-CL*, samples within the visual modality (green) and text modality (red) are clustered more closely together within each modality, and the distinction between samples across modalities is significantly enhanced, especially for the visual modality. Similarly, compared with subfigure (c), after applying the *VA-CL* algorithm, samples of the visual modality are more tightly clustered together in subfigure (d), and exhibit a clear differentiation from the samples of the acoustic modality (blue). This illustrates that *VA-CL* captures the unique emotional states of each modality as well as the emotional correlations between different samples.

Visualization of VA-Driven Emotion Correlation Prior. In Figure 4, we show the derived correlation matrices between emotions, i.e., **M** in Eqn. (8) on the two datasets. For instance, emotions such as *happy* and *surprise*, exhibit a relatively high positive correlation, while *happy* and *sad*, along with *neutral* and *surprise*, show a significantly high negative correlation. This aligns with our commonsense understanding of these emotional relationships.

Sensitivity Study of Threshold δ. Hyper-parameter δ determines whether the sample pair is positive or negative. As shown in the Figure 5, our UniVA achieves the best performance when δ is set to 0.1; moreover, it is observed that at values of 0.05 and 0.15, the model's performance is approximately the same; additionally, the performance of UniVA gradually decreases as the value of δ increases beyond 0.15.

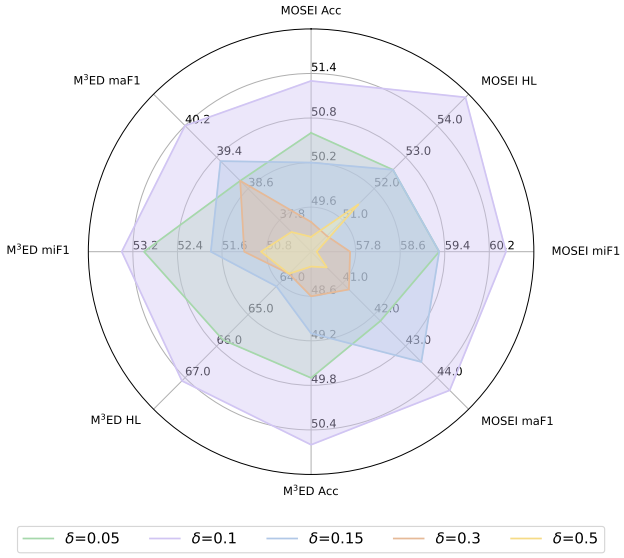


Figure 5: Sensitivity study of hyper-parameter δ. To better illustrate the results, we have taken the reciprocal of the HL metric and magnified it by 10 times.

4.6 Case Study

To better demonstrate the reasonability of the obtained VA scores for each modality, we present two test examples along with predictions from different methods. In Table 7 (a), due to the complexity of the ground-truth emotion labels, both *TAILOR* and *FacialMMT* missed one emotion and gave the incorrect prediction; for example (b), the emotional tendency displayed by the textual modality is exceedingly apparent, which results in that *TAILOR* only accurately predicted the dominant emotion reflected by the text. In both cases, our *uniVA* correctly classified the multi-label emotion categories, which shows the advantage of our framework by leveraging the VA scores of each modality.

5 Conclusion

In this paper, we proposed a **Uni**modal **V**alence-**A**rousal driven contrastive learning framework (UniVA) for the MMER task. Specifically, UniVA employs pre-trained VA models to obtain VA scores for each modality of all training samples, which are used to construct positive and negative samples for contrastive learning to obtain VA-aware unimodal representations. UniVA then integrates the unimodal representations and incorporates the emotion correlation prior in the VA space for emotion prediction. Experimental results on two datasets show the effectiveness of our UniVA model.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by the Natural Science Foundation of China (No. 62076133, 62006117, and 62272232).

References

[1] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of AACL*.

[2] Shahin Amiriparian, Lukas Christ, Andreas König, Alan S. Cowen, Eva-Maria Messner, Erik Cambria, and Björn W. Schuller. 2023. MuSe 2023 Challenge: Multimodal Prediction of Mimicked Emotions, Cross-Cultural Humour, and Personalised Recognition of Affects. In *Proceedings of ACM MM*.

[3] Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation. In *Proceedings of ACM MM*.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

[5] Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review* (2006).

[6] Sven Buechel and Udo Hahn. 2017. EMOBANK: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of EACL*.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* (2008).

[8] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*.

[9] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. In *Proceedings of EMNLP*.

[10] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979* (2020).

[11] Jiawen Deng and Fuji Ren. 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing* (2020).

[12] Hao Fei, Donghong Ji, Yue Zhang, and Yafeng Ren. 2020. Topic-enhanced capsule network for multi-label emotion classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2020).

[13] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of AAAI*.

[14] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* (2023).

[15] Shiping Ge, Zhiwei Jiang, Zifeng Cheng, Cong Wang, Yafeng Yin, and Qing Gu. 2023. Learning Robust Multi-Modal Representation for Multi-Label Emotion Recognition via Adversarial Masking and Perturbation. In *Proceedings of the ACM Web Conference*.

[16] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Proceedings of EMNLP Findings*.

[17] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of EMNLP*.

[18] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of ACM MM*.

[19] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, Wei Dang, and Xiaoying Pan. 2021. Deep Learning for Depression Recognition with Audiovisual Cues: A Review. *ArXiv* (2021).

[20] Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Song Hu. 2023. Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations. In *Proceedings of ACL*.

[21] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of EMNLP*.

[22] 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*.

[23] Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar R Zaiane. 2021. Seq2Emo: A sequence to multi-label emotion classification model. In *Proceedings of NAACL*.

[24] Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of ACM MM*.

[25] Aaron Keesing, Yun Sing Koh, Vithya Yogarajan, and Michael Witbrock. 2023. Emotion Recognition ToolKit (ERTK): Standardising Tools For Emotion Recognition Research. In *Proceedings of ACM MM*.

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of NeurIPS*.

[27] Dimitrios D. Kollias, Panagiotis Tzirakis, Alice Baird, Alan S. Cowen, and Stefanos Zafeiriou. 2023. ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[28] Solomon Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* (1951).

[29] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911* (2023).

[30] Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *Proceedings of AAAI*.

[31] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. GCNet: graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[32] Zheng Lian, Haiyang Sun, Licai Sun, Jinming Zhao, Ye Liu, B. Liu, Jiangyan Yi, Meng Wang, E. Cambria, Guoying Zhao, Björn Schuller, and Jianhua Tao. 2023. MER 2023: Multi-label Learning, Modality Robustness, and Semi-Supervised Learning. In *Proceedings of ACM MM*.

[33] 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* (2019).

[34] Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024. EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. *arXiv preprint arXiv:2401.08508* (2024).

[35] Reza Lotfian and Carlos Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Transactions on Affective Computing* (2019).

[36] Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of COLING*.

[37] Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of ACL*.

[38] 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*.

[39] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of AAAI*.

[40] Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of ACL*.

[41] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* (2017).

[42] Tongjie Pan, Yalan Ye, Hecheng Cai, Shudong Huang, Yang Yang, and Guoqing Wang. 2023. Multimodal Physiological Signals Fusion for Online Emotion Recognition. In *Proceedings of ACM MM*.

[43] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Heeyoung Park, and Alice H. Oh. 2019. Dimensional Emotion Detection from Categorical Emotion. In *Proceedings of EMNLP*.

[44] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

[45] Fan Qi, Zixin Zhang, Xianshan Yang, Huaiwen Zhang, and Changsheng Xu. 2022. Feeling Without Sharing: A Federated Video Emotion Recognition Framework Via Privacy-Agnostic Hybrid Aggregation. In *Proceedings of ACM MM*.

[46] Björn Schuller, Michel F. Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of International Conference on Multimodal Interaction*.

[47] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of AAAI*.

[48] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of ACL*.

[49] Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. 2023. How You Feelin'? Learning Emotions and Mental States in Movie Scenes. In *Proceedings of CVPR*.

[50] 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*.

[51] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[52] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* (2021).

[53] 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*.

[54] Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023. An Empirical Study on Multiple Knowledge from ChatGPT for Emotion Recognition in Conversations. In *Proceedings of EMNLP Findings*.

[55] Michel F. Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*.

[56] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* (2008).

[57] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[58] Xiangyu Wang and Chengqing Zong. 2021. Distributed representations of emotion categories in emotion space. In *Proceedings of ACL*.

[59] Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-Level Contrastive Learning for Emotion Recognition in Conversations. *IEEE Transactions on Affective Computing* (2023).

[60] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of ACL*.

[61] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of AAAI*.

[62] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of ACL*.

[63] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of EMNLP*.

[64] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of AAAI*.

[65] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of AAAI*.

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

[67] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT Equipped with Emotional Dialogue Capabilities? *arXiv preprint arXiv:2304.09582* (2023).

[68] Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations. In *Proceedings of ACL*.

[69] Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *Proceedings of EMNLP*.

[70] Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating Multi-turn Emotional Support Conversation with Positive Emotion Elicitation: A Reinforcement Learning Approach. In *Proceedings of ACL*.

[71] Yangyang Zhou, Xin Kang, and Fuji Ren. 2023. Prompt Consistency for Multi-label Textual Emotion Detection. *IEEE Transactions on Affective Computing* (2023).

[72] Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial Attention Modeling for Multi-dimensional Emotion Regression. In *Proceedings of ACL*.