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

Wenjie Zheng
Nanjing University of Science and Technology

Oct.29 2024

Content

- Task Definition
- Motivation
- Solution
- Experimental Results

Task Definition

Multimodal Multi-label Emotion Recognition (MMER)

- Given a monologue or dialogue containing multiple utterances, identify one or more emotions for each utterance.

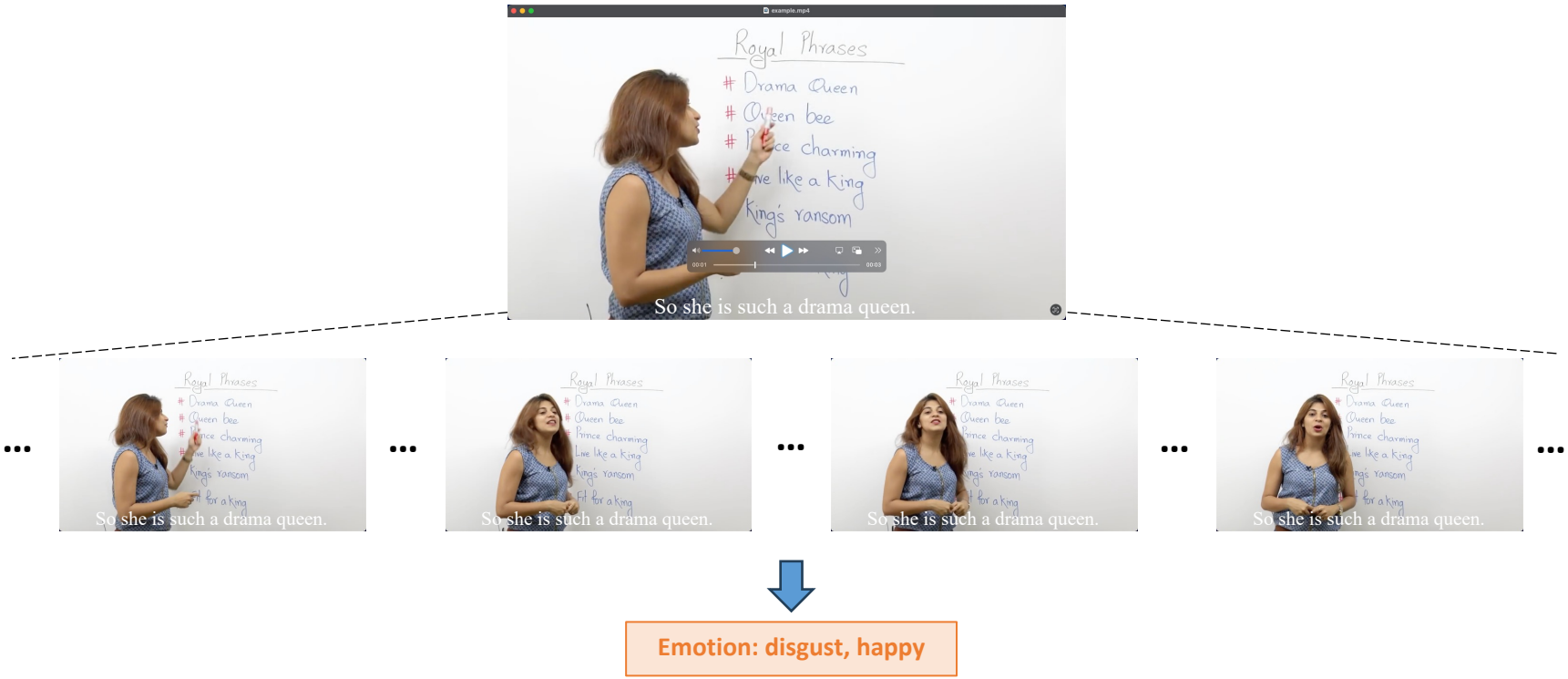


Figure 1: An example of the MMER task.

Motivation

- Previous studies represent each emotion category using a one-hot vector.
- Existing works mainly learn the unimodal representation based on the multimodal supervision signal of a single sample.

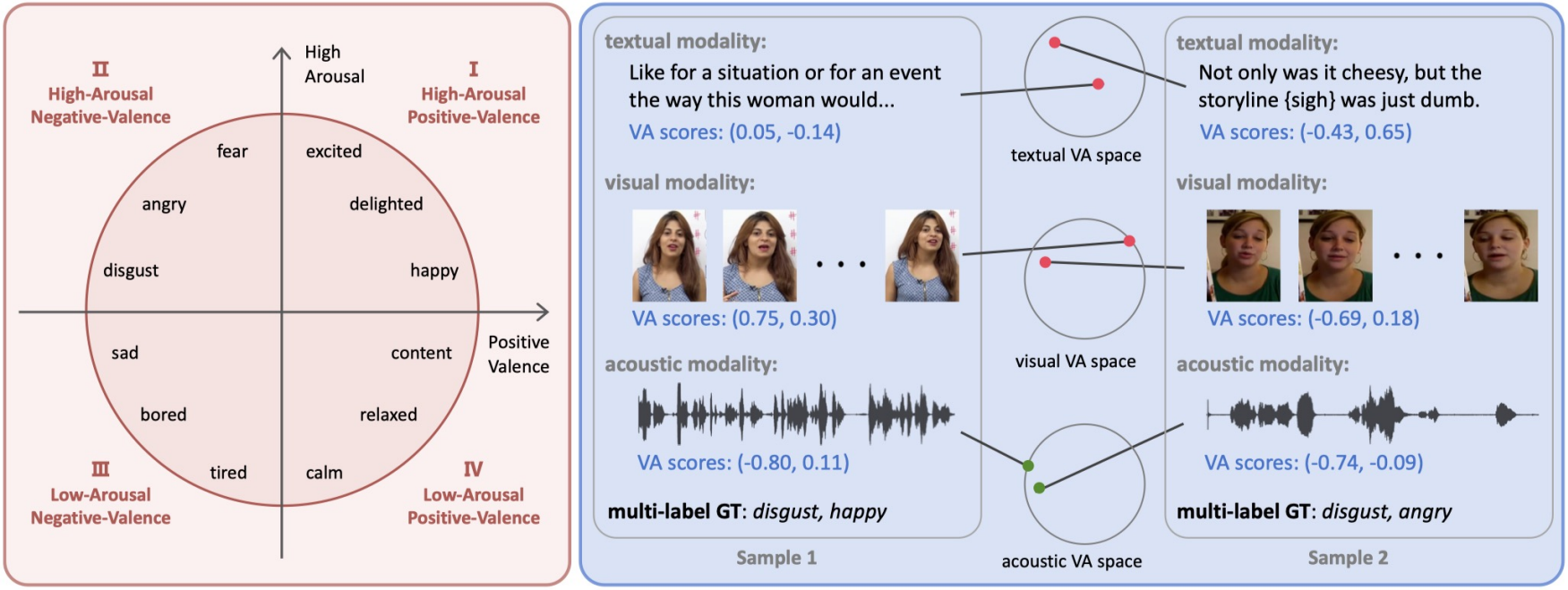


Figure 2: The left figure represents a **V**alence-**A**rousal (continuous dimensional) emotion space, while the right shows two examples from the MMER task.

Framework

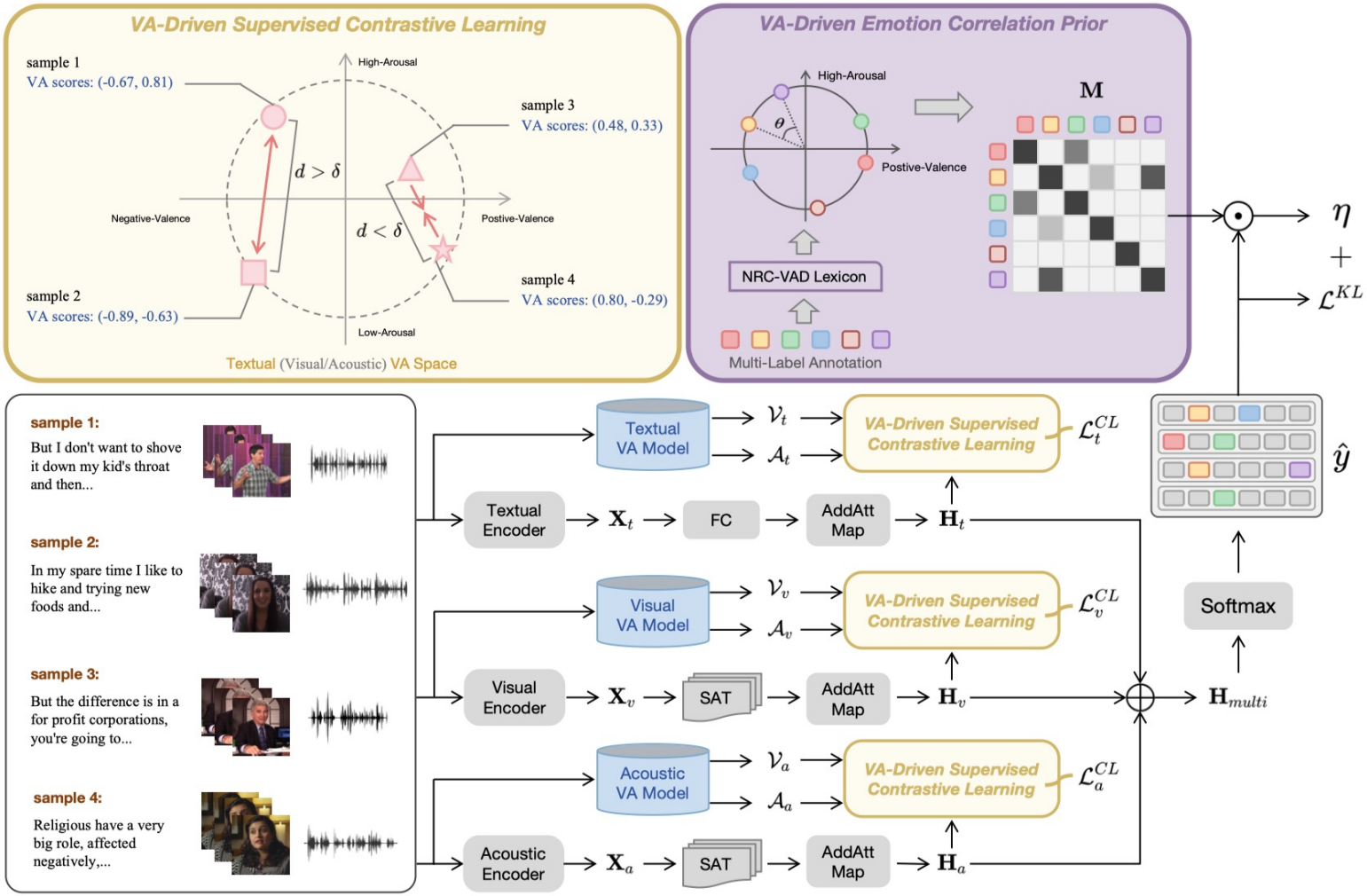
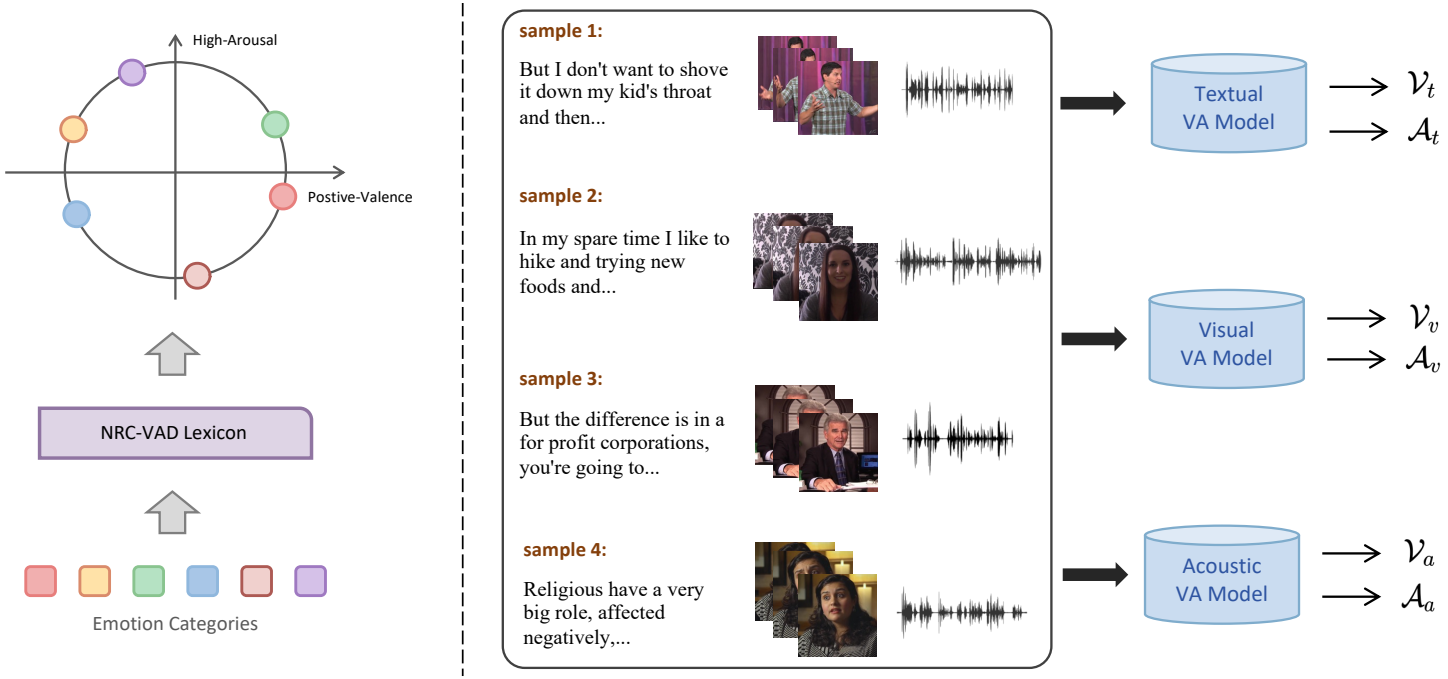


Figure 3: The overview of our proposed Unimodal Valence-Arousal driven contrastive learning framework (**UniVA**).

Module 1: VA Scores Acquisition

- **Label**: Utilize the NRC-VAD lexicon.
- **Each Modality**: Inference on different pre-trained VA models.



南京理工大学 文本挖掘实验室
Text Mining Lab, Nanjing University of Science & Technology

Module 2: VA-Driven Contrastive Learning for Each Modality

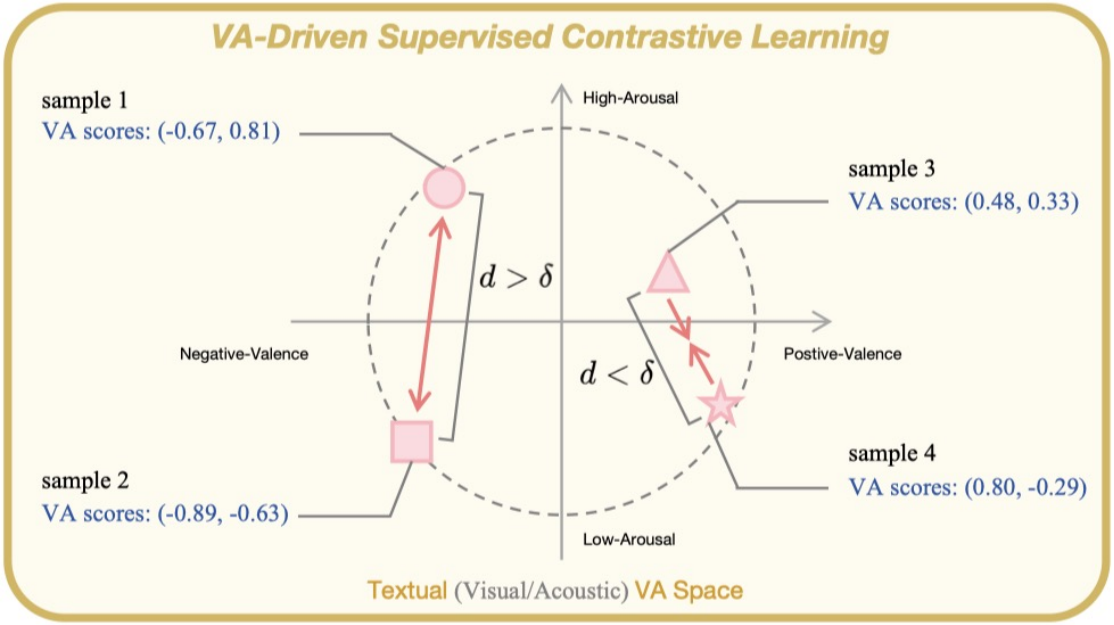
- Utilize the VA scores of each modality to construct the positive and negative samples:

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

- Calculate the loss of contrastive learning:

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

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



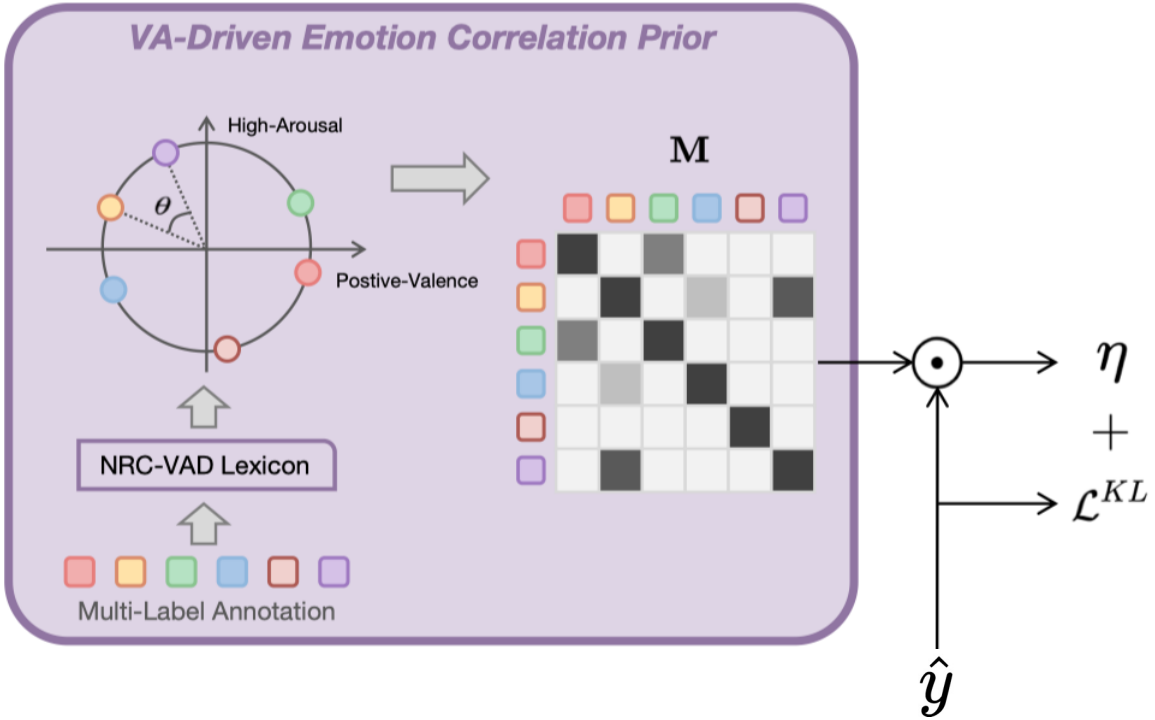
Module 3: VA-Driven Emotion Correlation Prior

- Calculate the emotion similarity matrix:

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

- Calculate the emotion correlation prior:

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



Experimental Results

- Main Results

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	43.0	-	-	-	-
TAILOR♠ (Zhang et al. [65])	43.7	0.206	49.7	37.1	-	-	-	-
RobMMR♠ (Ge et al. [15])	48.4	0.185	56.9	41.7	-	-	-	-
MDI♣ (Zhao et al. [66])	49.9	0.186	50.2	10.9	47.6	0.159	51.9	33.6
FacialMMT♣ (Zheng et al. [68])	50.1	0.190	59.1	40.8	48.7	0.154	51.7	37.9
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**

Table 1: Comparison with previous methods on benchmark datasets MOSEI and M^{3}ED.

Experimental Results

- Ablation Study

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

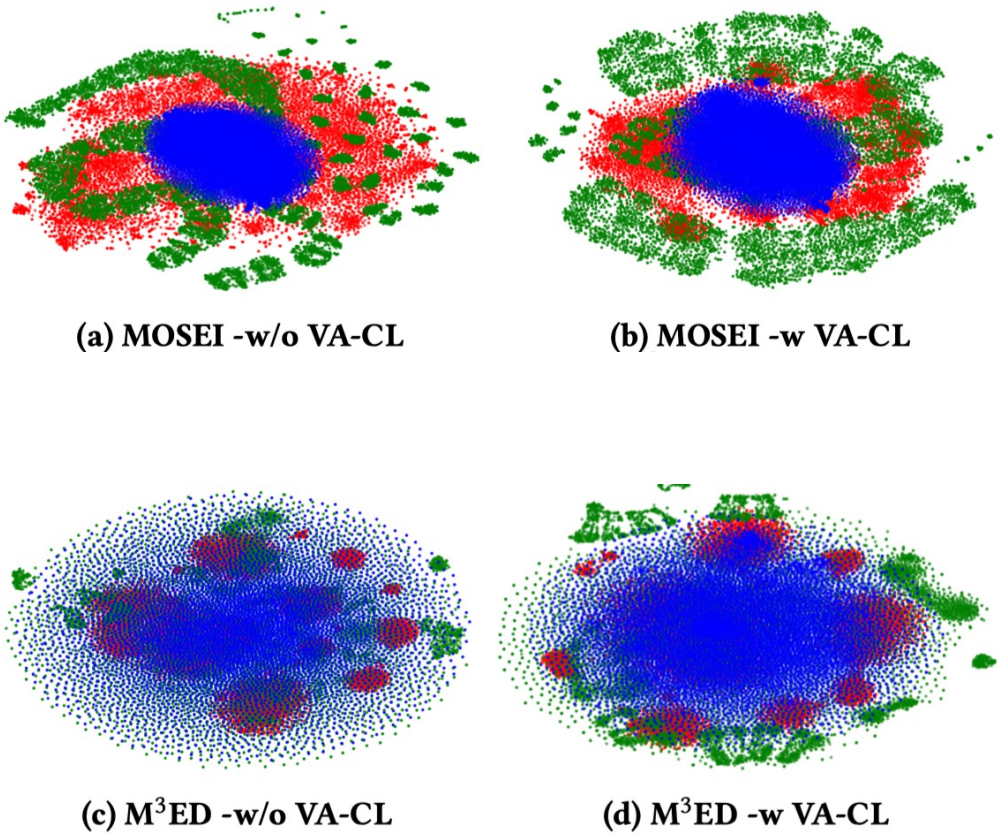
Table 2: Ablation study of our UniVA framework. VA-CL denotes VA-Driven contrastive learning, and VA-ECP denote VA-Driven emotion correlation prior.

Methods	MOSEI		M^{3}ED	
	Acc	*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 3: Ablation study of UniVA on different modalities.

Experimental Results

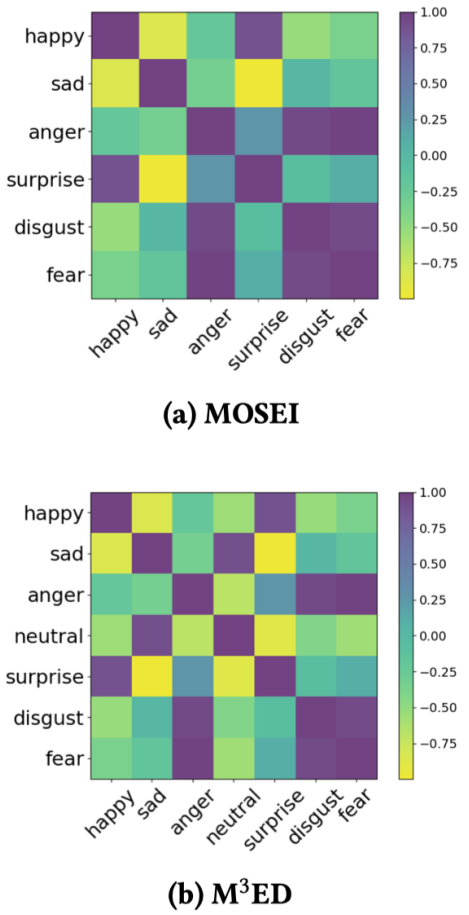
- Visualization



(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 4: 2D visualization of each modality on the training set.



(a) MOSEI

(b) M^{3}ED

Figure 5: The heatmap of VA-driven emotion correlation matrix.

Experimental Results

- Case Study

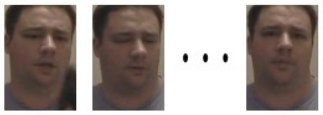
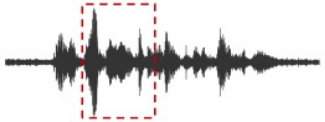
	(a)	(b)	(c)	(d)
Textual Modality	(umm) Just some untasteful things in the movie. (disgust and a little anger)	I would definitely recommend this, like I said, it's one of the classics. (happy)	咱妈会这么无聊吗? (Would our mom be so bored?) (disgust)	这叫居心叵测 (It's called an ulterior motive.) (angry and disgust)
Visual Modality	 disgust face … sad face	 sad face … sad face	 neutral face … disgust face	 angry face, angry face … disgust face
Acoustic Modality	 angry voice	 happy voice	 neutral voice, disgust voice	 angry voice
GT	(*disgust, angry, sad*)	(*happy, sad*)	(*neutral, disgust*)	(*angry, disgust*)
TAILOR	(disgust, angry) ✗	(happy, sad) ✓	(disgust) ✗	(angry, disgust) ✓
FacialMMT	(disgust, sad) ✗	(happy) ✗	(neutral, disgust) ✓	(angry, disgust) ✓
UniVA	$(VA)_{textual}$ Scores: (-0.43, 0.08) $(VA)_{visual}$ Scores: (-0.36, -0.02) $(VA)_{acoustic}$ Scores: (-0.67, 0.51) (disgust, angry, sad) ✓	$(VA)_{textual}$ Scores: (0.80, 0.55) $(VA)_{visual}$ Scores: (-0.62, -0.27) $(VA)_{acoustic}$ Scores: (0.74, 0.59) (happy, sad) ✓	$(VA)_{textual}$ Scores: (-0.17, -0.33) $(VA)_{visual}$ Scores: (-0.09, -0.28) $(VA)_{acoustic}$ Scores: (-0.23, -0.31) (neutral, disgust) ✓	$(VA)_{textual}$ Scores: (-0.26, -0.43) $(VA)_{visual}$ Scores: (-0.45, -0.47) $(VA)_{acoustic}$ Scores: (-0.64, -0.72) (angry, disgust) ✓

Table 4: Prediction comparison on two samples from the test sets of MOSEI and M^3ED.