

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

Anonymous Authors

1 DESIGN OF GEMINI'S PROMPTS

In this work, we select the current mainstream multimodal large language model, Gemini, as one of our baselines and conduct a zero-shot experiment. In our implementation, we designed two prompts: one for generating video caption, and another for generating the final multi-label emotion predictions. Specifically, the prompt for generating `<video_caption>` is: *Here are some frames from a video. Could you explain what the video is describing?*, and the prompt for generating the emotion prediction is: *Based on the following video description and the individual's monologue, please identify the emotions expressed. Video description is `<video_caption>`. Monologue is `<speech_content>`. Analyze and categorize the predominant emotion or emotions into one or more of these categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. Here, `<speech_content>` represents the spoken content in the current video.*

2 DETAILED ABLATION STUDY RESULTS

In this section, we present the results of ablation studies for the effect of each modality and the effect of different contrastive learning in detail, as shown in Tables 1, 2, and 3. Here, *VA-CL* indicates VA-driven unimodal contrastive learning, *Sup-CL* indicates supervised contrastive learning, and *Self-CL* indicates self-supervised contrastive learning.

Table 1: Ablation study of UniVA on different modalities for MOSEI.

Methods	MOSEI			
	Acc	HL	miF1	maF1
UniVA	51.4	0.185	60.1	43.5
- w/o Vision	51.1	0.186	59.9	42.4
- w/o Audio	49.7	0.187	58.1	38.2
- w/o Vision, Audio	50.6	0.190	59.4	42.8
- w/o Text, Vision	46.7	0.215	54.5	33.3
- w/o Text, Audio	42.3	0.235	48.2	23.9

3 ERROR ANALYSIS

In Table 4, we present an error analysis of our method *UniVA* by comparing its predictions for the MMER task with two baselines, *TAILOR* and *FacialMMT*, on two test examples. Specifically, in case (a), all three methods failed to recognize the *angry* emotion intended by the visual modality. This may be due to the emotion displayed in the visual modality being subtle and only appearing in a few video frames. In case (b), since the *disgust* emotion expressed in the textual modality was too strong, all three models only correctly predicted one.

Table 2: Ablation study of UniVA on different modalities for M³ED.





Methods	M ³ ED			
	Acc	HL	miF1	maF1
UniVA	50.6	0.149	53.4	40.2
- w/o Vision	49.2	0.156	52.0	40.0
- w/o Audio	48.4	0.157	51.4	38.9
- w/o Vision, Audio	48.0	0.154	51.6	37.8
- w/o Text, Vision	38.8	0.164	41.3	14.0
- w/o Text, Audio	40.8	0.188	40.7	11.1

Table 3: Ablation study of UniVA with different contrastive learning algorithm. The "-r" indicates that the proposed VA-CL is replaced with other algorithms.

Methods	MOSEI			
	Acc	HL	miF1	maF1
UniVA (VA-CL)	51.4	0.185	60.1	43.5
-r Sup-CL	50.3	0.190	59.0	41.4
-r Self-CL	49.6	0.193	58.2	40.8

Methods	M ³ ED			
	Acc	HL	miF1	maF1
UniVA (VA-CL)	50.6	0.149	53.4	40.2
-r Sup-CL	48.8	0.154	52.7	38.0
-r Self-CL	46.2	0.158	50.1	37.5

Table 4: Prediction comparison on the MOSEI and M³ED.

	(a)	(b)
Textual Modality	They are at one level but if you can do without... <i>neutral</i>	他说你就信? (Are you just going to believe him?) <i>disgust and a little angry</i>
Visual Modality	 <i>angry face</i>	 <i>neutral</i>
Acoustic Modality	 <i>sad voice</i>	 <i>a little angry</i>
GT	(sad, angry)	(disgust, angry)
TAILOR	(sad) ×	(disgust) ×
FacialMMT	(sad) ×	(disgust) ×
UniVA	(VA) _{textual} Scores: (-0.08, 0.04) (VA) _{visual} Scores: (-0.14, 0.11) (VA) _{acoustic} Scores: (-0.75, -0.22) (sad, neutral) ×	(VA) _{textual} Scores: (-0.89, 0.16) (VA) _{visual} Scores: (-0.12, 0.15) (VA) _{acoustic} Scores: (-0.15, 0.11) (disgust) ×