Supplementary Materials: TGCA-PVT for Sticker Emotion Recognition

1 THE SIGNIFICANCE OF RESEARCHING STICKER EMOTION RECOGNITION.

Stickers, as powerful carriers of images and text, often fulfill the role of "a picture worth a thousand words" in online chats, effectively conveying users' emotions as we can see in Fig. 1. In the same online conversation, while text alone may fail to accurately reflect emotions, a sticker can bridge this gap. We also make a comparison between general images and stickers in Fig. 1. Compared to recognizing sentiments in realistic images, identifying emotions in stickers is more challenging. Stickers represent multi-modal data, combining images and text, requiring models to understand both global and local information. Additionally, stickers can consist of general images, as well as various forms such as cartoons and drawings. Compared to the general image, the wide variety of topics reflected in stickers introduces significant domain differences, further testing the model's generalization capabilities.

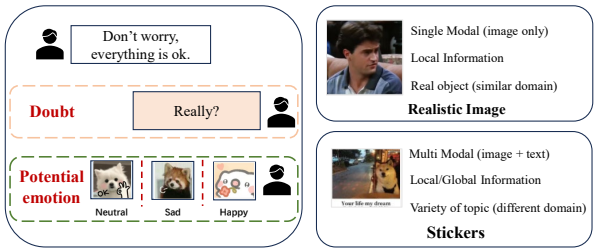

Figure 1: The significance of researching sticker emotion recognition.

2 ABOUT THE TOPIC ID.

The topic ID is merely a flag for grouping images by theme or origin, and its value is not meaningful for the topic feature. We use a pre-trained Bert tokenizer to get the word ID of the sticker theme, then convert it to a unique topic ID using hash mapping. For the FI dataset without themes, we use filenames to obtain the word ID. We can assign a corresponding ID to each image in scenarios without predefined image topics. Then in our method, this ID is used to group the stickers to consider the situation in Figure 1 (a) and (b) in our paper. By making the model extract related features of stickers with the same topic ID during the training process, the model will be able to better recognize the emotion because stickers with the same ID in the same batch will provide relevant features to each other.

3 MORE METRICS ABOUT THE EMOTION RECOGNITION ON THE SER30K DATASET.

We provided more classification metrics to evaluate better our proposed method, such as recall, precision, map, etc. The details can be seen in Table 1. Specifically, we also compared the latest image-based emotion recognition and large language models.

For the image-based emotion recognition method, we use the MAM that incorporates different visual concepts for emotion analysis [3]. Since MEM cannot deal with multi-modal data including text, we adopt two versions. The first one is the original MEM that discards the sticker text data. Then to deal with the text method, we use the same pre-trained Bert model to encode the text data. Then we flatten the text feature from the Bert model and concatenate it with the image feature for sticker emotion recognition.

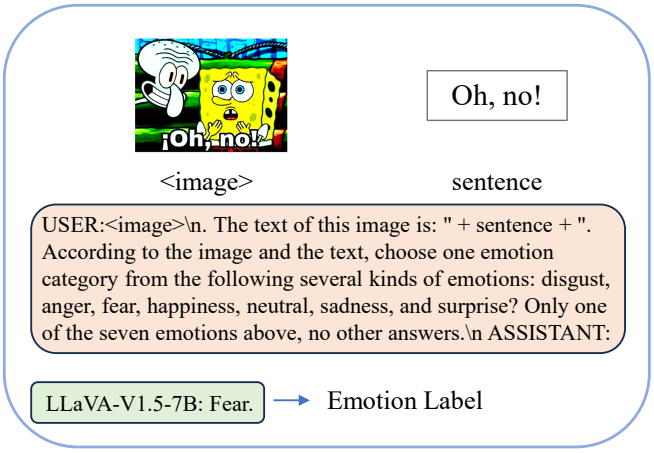

Figure 2: Framework of Using LLaVA for sticker emotion recognition.

Since multi-modal large language models, such as LLaVA and GPT4-V, have shown great performance in dealing with image-text data, we adopt the open-source LLaVA-v1.5-7b [1] model to conduct emotion recognition on the SER30K dataset as a comparison. We use simple instruction input to let the LLaVa model judge the emotion represented by the sticker, a case is shown in Fig. 2.

We directly load the offline LLaVA-v1.5-7b model, and then feed the sticker data in the test set with the corresponding text data to this model as in Figure 1, and prompt it to return the corresponding emotion category. During this process, we found that a small percentage of the time LLaVA returned non-sentiment text, which we treated as misclassification.

4 MORE DETAILS ABOUT THE ABLATION STUDY.

In sticker emotion recognition tasks, local information like expressions and poses is important. In addition, since some stickers have multiple subjects, it is also extremely critical to consider the global information relations among objects. To better verify the effect of considering both global and local features, we conduct experiments to remove global features and local features respectively as Table **??**.

Table 1: Comparison with Baselines on SER30K

	MAM			MAM + Bert			LORA			LLaVA-v1.5-7b			**TGCA-PVT**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Anger	55.67	34.72	42.78	60.82	37.82	46.64	60.33	46.18	52.32	57.32	49.82	53.31	65.67	43.82	**52.56**
Disgust	0.00	0.00	0.00	1.00	4.76	9.09	50.00	11.90	19.23	4.35	9.52	5.97	35.73	11.90	**17.86**
Fear	71.88	27.88	40.17	65.28	28.48	39.66	65.09	41.82	50.92	27.06	13.94	18.40	66.09	46.06	**54.29**
Happiness	78.50	82.72	80.55	79.09	82.32	80.67	76.83	83.83	80.18	75.23	72.72	73.96	79.57	82.72	**81.12**
Neutral	65.26	77.81	70.98	65.98	78.18	71.56	68.80	75.13	71.82	98.40	5.69	10.75	69.39	77.25	**73.11**
Sadness	61.27	50.15	55.16	61.20	54.46	57.64	67.70	58.33	62.67	27.15	84.52	41.10	63.62	62.20	**62.90**
Surprise	50.42	39.67	44.40	53.51	40.00	45.78	55.14	38.69	45.47	19.67	65.57	30.26	53.03	42.95	**47.46**
Macro avg	54.72	44.71	47.72	**69.41**	46.58	50.15	63.41	50.84	54.66	38.65	37.72	29.22	61.87	**52.42**	**55.61**
Weighted avg	67.81	68.97	67.60	69.39	69.76	68.58	68.96	70.75	69.91	**71.99**	46.02	42.16	71.03	**71.63**	**70.93**
MAP / Accuracy	44.01	-	68.97	45.92	-	69.75	55.12	-	70.75	-	-	46.02	**55.63**	-	**71.63**
Param/ Flops ($10\hat{6}$)	**26.85**	-	**165.57**	112.43	-	1910.56	142.33	-	179.96	-	-	-	144.35	-	201.79

Table 2: Additional Ablation Study on SER30K

Model	PVT+Bert			Without Global			Without Local			With LORA			With LERA		
TGCA-PVT	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Anger	59.18	45.45	51.67	59.52	44.91	51.19	55.00	52.00	53.45	60.54	44.91	51.57	65.67	43.82	52.56
Disgust	33.33	4.76	8.33	85.71	14.29	24.49	46.15	14.29	21.82	40.00	9.52	15.38	35.73	11.90	17.86
Fear	64.22	42.42	51.09	61.79	46.06	52.78	75.00	43.64	55.17	69.09	46.06	55.27	66.09	46.06	54.29
Happiness	77.93	82.23	80.25	78.27	82.90	80.52	78.72	82.19	80.42	77.52	83.79	80.53	79.57	82.72	81.12
Neutral	68.54	75.73	71.95	68.37	76.05	72.01	69.38	76.38	72.71	68.49	75.27	71.72	69.39	77.25	73.11
Sadness	63.16	59.67	61.84	68.26	57.29	62.30	67.50	55.95	61.18	66.67	58.63	62.39	63.62	62.20	62.90
Surprise	50.43	38.36	43.58	52.21	42.62	46.93	51.08	38.39	44.03	53.88	38.69	45.04	53.03	42.95	47.46
Macro avg	59.82	50.09	52.86	67.73	52.02	55.54	63.26	51.87	55.54	62.31	50.98	54.55	61.87	52.42	55.61
Weighted avg	68.77	69.59	68.87	70.33	70.84	70.12	70.39	70.92	70.31	69.98	70.80	69.97	71.03	71.63	70.93
MAP / Accuracy	51.62	-	68.35	54.06	-	70.81	55.23	-	70.93	53.32	-	70.80	55.63	-	71.63

Specifically, we set up four variants of the model: 'PVT+Bert', which is the combination of the basic skeleton network PVT and Bert model; 'Without Global' means removing the global features of visual modalities X_{global} extracted by TGVA-PVT. 'Without Local' means removing the local features of visual modalities X_{local} extracted by TGVA-PVT. 'With LORA' means that the LORA mechanism proposed in Ref. [2] is used instead of the proposed LERA mechanism. The final one called 'With LERA' is our proposed TGCA-PVT model.

5 SENTIMENT DISTRIBUTION.

To better see how the classification behaves by TGCA-PVT on the SER30K dataset, we conduct a visualization of the sentiment distribution. Specifically, we used TGCA-PVT to encode the stickers and obtained the hidden encoding before the classify head. Then we use Linear Discriminant Analysis (LDA) to reduce the obtained latent representations to a 2-dimensional space, and visualize them with different colors based on their true labels, the results are shown in Fig. 3.

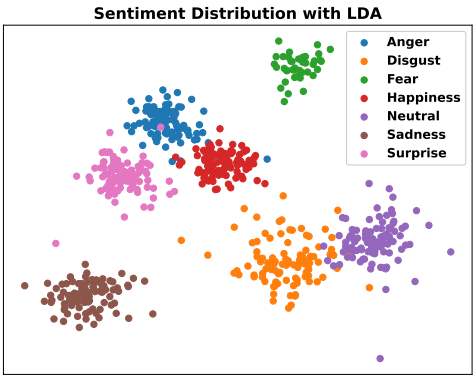

Figure 3: Sentiment Distribution by TGCA-PVT on the SER30K dataset.

REFERENCES

[1] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[2] Shengzhe Liu, Xin Zhang, and Jufeng Yang. 2022. SER30K: A large-scale dataset for sticker emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 33–41.

[3] Hao Zhang, Gaifang Luo, Yingying Yue, Kangjian He, and Dan Xu. 2024. Affective image recognition with multi-attribute knowledge in deep neural networks. *Multimedia Tools and Applications* 83, 6 (2024), 18353–18379.