

A SUPPLEMENTARY MATERIAL

A.1 Details of our compared methods

The methods compared in our experiments are all the latest state-of-the-art methods. FT[4] is published in IEEE TMM 2022, and it is a pioneering work for emotional video captioning, which firstly adds emotional cues to caption generation. They design two independent prediction networks: fact-stream and emotion-stream. In addition, they release the first emotional video captioning dataset, which has made a huge contribution to research in this task. CANet[2] is published in IEEE TMM 2023, and it proposes a contextual attention network for EVC task, which leverage semantic-rich context learning to recognize and describe the fact and emotion in the video. EPAN[1] is published in ACM MM 2023, and it proposes a tree-structured emotion learning module, which utilizes hierarchical emotion trees to learn emotion categories and emotion words, generating more explicit emotional cues. VEIN[3] is published in IEEE TIP 2024, and it proposes a vision-based emotion interpretation network, which incorporate visual context, textual context, and visual-textual relevance into an aggregated multimodal contextual vector to enhance video captioning.

Compared with these state-of-the-art methods, our method is the first work to consider to model the dynamic changes of emotion and adaptively generate emotion-related words at the necessary generation steps, making our captions vivid and diverse.

A.2 Additional Experimental Results

Effect of Different Evolution Orders. In this work, we first leverage element-level evolution to refine the dynamic factor for subspace evolution. To explore the impact of evolution orders for the results, we try two different orders of emotion evolution, including “subspace-level -> element-level” and “element-level -> subspace-level”. As shown in Table 1, the second evolution order, which we adopt in the paper, brings better performance, i.e., 1.3% and 1.9% on $Acc_{S,W}$ and ROUGE metrics than the first evolution order. We firstly perform global element-level evolution of emotion based on the generated dynamic evolution factors, without limiting the intensity and direction of evolution. It may evolve from any emotion to any other emotion. Element-level evolution avoid us from falling into the local optima. We then utilize subspace weighted recompose for fine-grained evolution. Thus, our evolution manner is consistent with the natural rules of feature evolution.

Effect of Different Subspace Lists. The subspace list is an important hyper-parameter that affects subspace-level emotion evolution. Different lists determine different subspace combinations and affect the direction of evolution. In Table 2, we show the performance of different subspace number list. We observe that when the list is set to {2,3,5,6,10}, our method achieve significant improvements in semantic and hybrid metrics, with only a slight lower in accuracy metrics, i.e., reach 28.1 and 47.7 on METEOR and ROUGE metrics. Firstly, from the results, we find that in subspace-level recompose evolution, too few subspaces will lead to incomplete semantic space recompose, thereby affecting subspace-level emotion evolution. Secondly, we find that if the subspace dimension is too small, the semantic information of emotional features will be lost. In

short, choosing an complete and suitable subspace number list can greatly improve the performance of the subspace re-composition.

Table 1: Ablation studies of different emotion evolution orders on the EVC-VE dataset. S and E denotes the subspace-level emotion re-composition and element-level emotion evolution, respectively.

Orders	$Acc_{S,W}$	Acc_C	B1	B2	B3	B4	M	R	C	BFS	CFS
S->E	70.1	68.6	73.3	54.9	39.7	27.8	22.8	46.8	41.0	46.5	46.2
E->S	71.0	69.4	74.5	55.3	40.0	28.1	23.4	47.7	41.5	47.3	46.9

Table 2: Ablation studies of subspace number list on the EVC-VE dataset.

List	$Acc_{S,W}$	Acc_C	B1	B2	B3	B4	M	R	C	BFS	CFS
{2,3,5}	70.2	68.4	74.3	54.8	39.1	27.5	22.6	47.1	40.5	46.5	45.8
{2,3,4,5,6}	70.6	69.1	74.3	55.2	39.5	27.5	22.8	47.0	41.0	47.0	46.8
{2,3,5,6,10}	71.0	69.4	74.5	55.3	40.0	28.1	23.4	47.7	41.5	47.3	46.9
{2,3,100,150,300}	69.8	67.8	74.0	54.6	38.9	27.1	22.4	46.7	39.2	45.8	44.5
{2,3,5,60,100,150,300}	71.1	69.6	73.9	54.3	39.0	26.9	22.5	46.8	40.3	46.6	45.8

REFERENCES

- [1] 2023 MM. Emotion-prior awareness network for emotional video captioning. 589–600.
- [2] Peipei Song, Dan Guo, Jun Cheng, and Meng Wang. 2022. Contextual attention network for emotional video captioning. *IEEE Transactions on Multimedia* (2022).
- [3] Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. 2024. Emotional Video Captioning with Vision-based Emotion Interpretation Network. *IEEE Transactions on Image Processing* (2024).
- [4] Hanli Wang, Pengjie Tang, Qinyu Li, and Meng Cheng. [n. d.]. Emotion expression with fact transfer for video description. *IEEE Transactions on Multimedia* ([n. d.]).