

Supplementary Materials: HESP

The content of our supplementary material is organized as follows.

- Fixed vs. Learnable Textual Prompting
- Various areas for Learnable Visual Representations
- Effects of Negative Representations
- Effects of Different Probability Scores
- Comparison of Various CLIPs
- Confusion Matrices on Four OV-FER Tasks
- Visualization on Various Features
- Feature Maps and Prediction Results of Different Methods

1 Various settings in HESP

1.1 Fixed vs. Learnable Textual Prompting

In order to explore the impact of different textual prompting (TP) methods with HSEP, we reported the results with various settings in TP in Table 1. For comparison, referring to the normal CLIP [4], we set fixed textual prompts as "the emotion of this video is {" for closed-set learning and "this video is not {" for open-set learning, respectively. The results show that the best performance is achieved when using the learnable textual prompt representations for closed-set data and fixed negative textual prompts for open-set data. Compared to the fixed textual prompts for both closed-set and open-set learning, our method achieves an increase of 4.55% on AUROC and 2.96% on OSCR, relatively. This proves that for closed-set data, learnable textual prompts are superior to fixed textual prompts, obtaining more information related to emotions during the training process. As for open-set data, since they are unknown during the training process, we need intuitive negative-category information to indirectly introduce open-set information, rather than exploring negative-category information through learnable textual prompts.

Table 1: Effects of various settings in the textual prompting module on OV-FER

For closed-set		For open-set		AUROC	OSCR
Fixed	Learnable	Fixed	Learnable		
✓		✓		61.59	40.53
✓			✓	60.19	40.18
	✓		✓	62.79	39.63
	✓	✓		64.39	41.73

1.2 Various areas for Learnable Visual Representations

In order to verify the effect of different face areas for learnable visual prompt representations, we divided the video frame into 16 facial areas based on a 4×4 grid, each of which can be used for learnable visual representations, respectively, with the size of 56×56 , as shown in Fig. 1(a). We conducted experimental evaluation for each area on the 7 basic emotion OV-FER task with the openness $O(2 : 5) = 0.47$. The results are shown in Fig. 1(b), where each

square represents the AUROC value at the corresponding facial location. Clearly, facial area 6 (eye area) offers the best results, followed by area 8 (eye corner position) and 16 (mouth area). It is evident that these facial areas are sensitive to expression. Thus, this further supports our conclusion that better results can be achieved by placing learnable visual representations in expression-sensitive areas, such as the eyes and corners of the mouth.

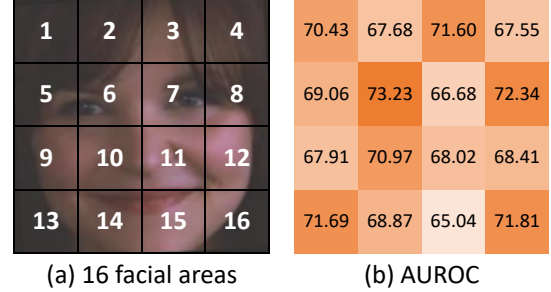


Figure 1: Different areas for learnable visual representations. The superior results are achieved when focusing on the area 6, specifically the expression-sensitive eye region, suggesting that leveraging the expression-sensitive area aids CLIP in capturing more potent emotion cues.

In addition, we also conducted the comparison experiments of expression-sensitive area selection for videos, namely in the first frame of a video vs. each frame of a video. The results are shown in Table 2. The results show that using the first frame to extract expression-sensitive areas achieves subtle improvement, indicating that keeping the rectangular area stable throughout the video is more beneficial.

Table 2: Effects of expression-sensitive area extraction in different frames of a video.

	AUROC	OSCR
each frame	62.54	41.59
first frame	64.39	41.73

1.3 Effects of Negative Representations

To further discuss the effects of negative representations in HESP, we conducted three different settings in the negative representation learning, including using only fixed negative textual representations (Fixed-NT), using only learnable negative textual representations (Learnable-NT), and using both Fixed-NT and learnable negative visual representations (Learnable-NV) in our paper. The comparison results are shown in Table 3. From the results, we can observe that using only Fixed-NT performs poorly, whereas utilizing only

Learable-NT achieves a relative improvement of 6.02% and 6.74% on AUROC and OSCR, respectively. This indicates that Learable-NT enable the model to learn some information about unknown classes. Combining the Fixed-NT with Learable-NV resulted in optimal performance (more details in our paper), leading to a relative improvement of 5.35% in AUROC and 13.12% in OSCR compared to the second setting. This indicates that while the Learable-NT assists in learning information about unknown classes, its performance might be affected by visual noise. Therefore, by leveraging the Learable-NV and Fixed-NT to provide intuitive negative class information, we can better introduce novel information about unknown classes.

Table 3: Effects of different settings in negative representations for HESP.

Different Settings	AUROC	OSCR
only Fixed-NT	57.65	34.56
only Learable-NT	61.12	36.89
Fixed-NT & Learable-NV	64.39	41.73

1.4 Effects of Different Probability Scores

When using our method for open-set facial expression prediction, three different probabilities are generated: known class prediction probability P_{KN} , negative representation prediction probability P_{NE} , and overall class prediction probability P_H . More details can be seen in Sec 3.5 of our paper. Table 4 reports the prediction results of using different probabilities for OV-FER. The results indicate a significant decrease in performance when using P_{KN} , showing that using only closed-set probability to predict open-set data is unreliable. When using P_{NE} , the AUROC and OSCR are relatively increased by 13.26% and 29.73% compared to P_{KN} , because the calculation process of P_{NE} introduces unknown classes information, which is significantly beneficial for open-set prediction. In the end, when using P_H , both evaluation metrics reached the optimal result, proving the necessity of integrating P_{KN} and P_{NE} for the OV-FER prediction.

Table 4: Comparison of different probability scores with HESP

Probability	AUROC	OSCR
P_{KN}	53.94	30.14
P_{NE}	61.09	39.10
P_H	64.39	41.73

1.5 Comparison of Various CLIPs

To verify the effectiveness of our proposed HESP for augmenting CLIP on OV-FER, we compared our HESP with three various CLIP methods, including vanilla CLIP[4], ViFi-CLIP[5], and Open-VCLIP[6], respectively, on the 7 basic emotion OV-FER task. The experimental results are shown in Table 5. Specifically, ViFi-CLIP transferred CLIP to the video domain based on text prompts and fine-tuning. Open-VCLIP extended the temporal attention view of each self attention layer in vanilla CLIP to facilitate the aggregation of global temporal information, and also used hand-crafted text prompts to assist visual encoders in learning features. The experimental results demonstrate that our HESP effectively augments various CLIPs for OV-FER, indicating the robustness and generalization of our HESP.

Table 5: Comparison of various CLIP methods on OV-FER

Various CLIPs	AUROC	OSCR
CLIP[4]	54.60	20.24
ViFi-CLIP[5]	55.98	34.38
Open-VCLIP[6]	56.53	24.47
OURS	64.39	41.73

2 VISUALIZATION

2.1 Confusion Matrices on Four OV-FER Tasks

Fig. 2-5 show the confusion matrices for our method on four different OV-FER tasks. For each task, we separately evaluate the performance of our method on closed-set and open-set data to demonstrate the accuracy of known class classification and unknown class recognition. Specifically, Fig. 2 and Fig. 3 show the confusion matrices of our method under four different openesses on the 7 basic emotion OV-FER task and the 11 emotion OV-FER task, respectively. We observe that, with fewer open-set data, we achieve higher unknown class recognition accuracy. Yet, as openness increases, open-set recognition accuracy declines. Conversely, reducing the number of known classes enhances closed-set classification accuracy. Fig. 4 and Fig. 5 illustrate the confusion matrices of our method on the fusion dataset OV-FER task and the compound emotion OV-FER task, respectively. These two tasks are confronted with a larger volume of data and a more complex open environment. The results indicate that our method can also effectively identify the unknown class and achieve high accuracy on challenging closed-set data, particularly for the facial expression categories of happy, neutral, and sad.

2.2 Visualization on Various Features

Fig. 6-9 illustrate high-level facial expression features extracted by ARPL [1], CLIP+ARPL, Open-VCLIP [6], and our HESP on four different OV-FER tasks, respectively. Obviously, compared to other methods, the features extracted by our HESP demonstrate the most significant feature separation effects across four OV-FER tasks. The feature separation is most effective in the 7 basic emotion OV-FER task due to its smaller number of categories, allowing the model

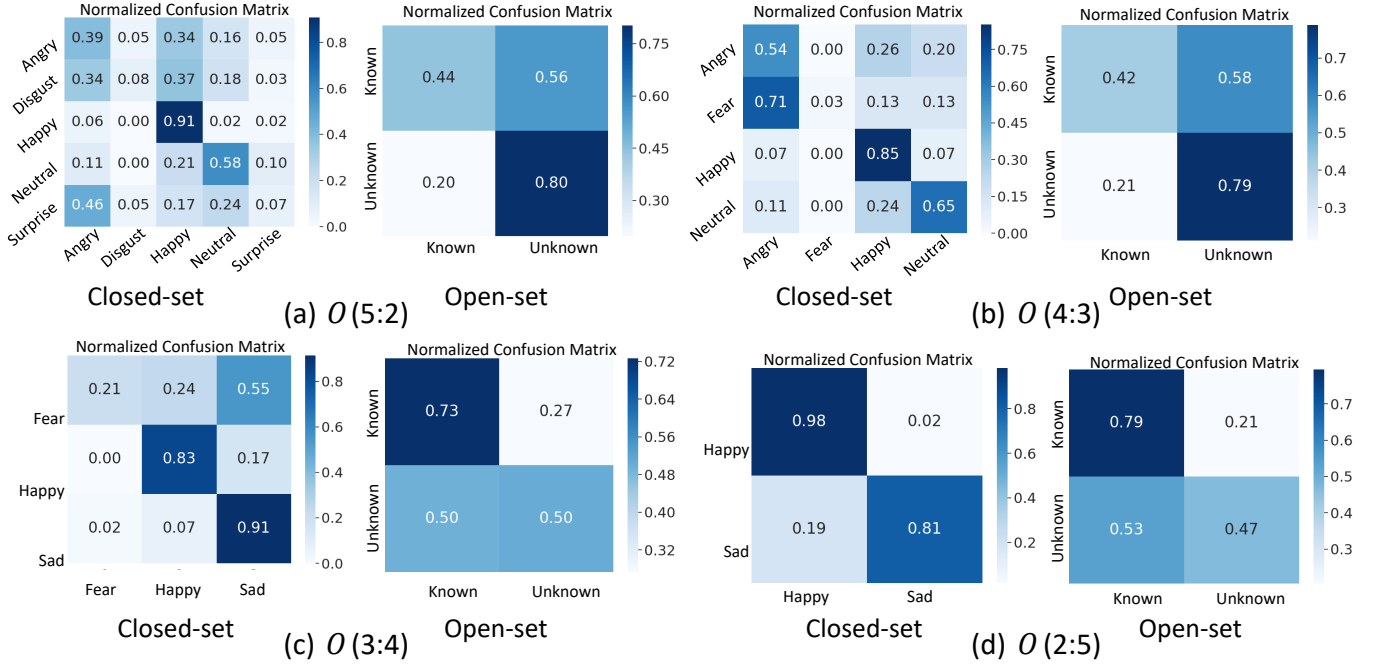


Figure 2: The confusion matrices of our method on the 7 basic emotion OV-FER task. Known and unknown classes are divided according to four different opennesses.

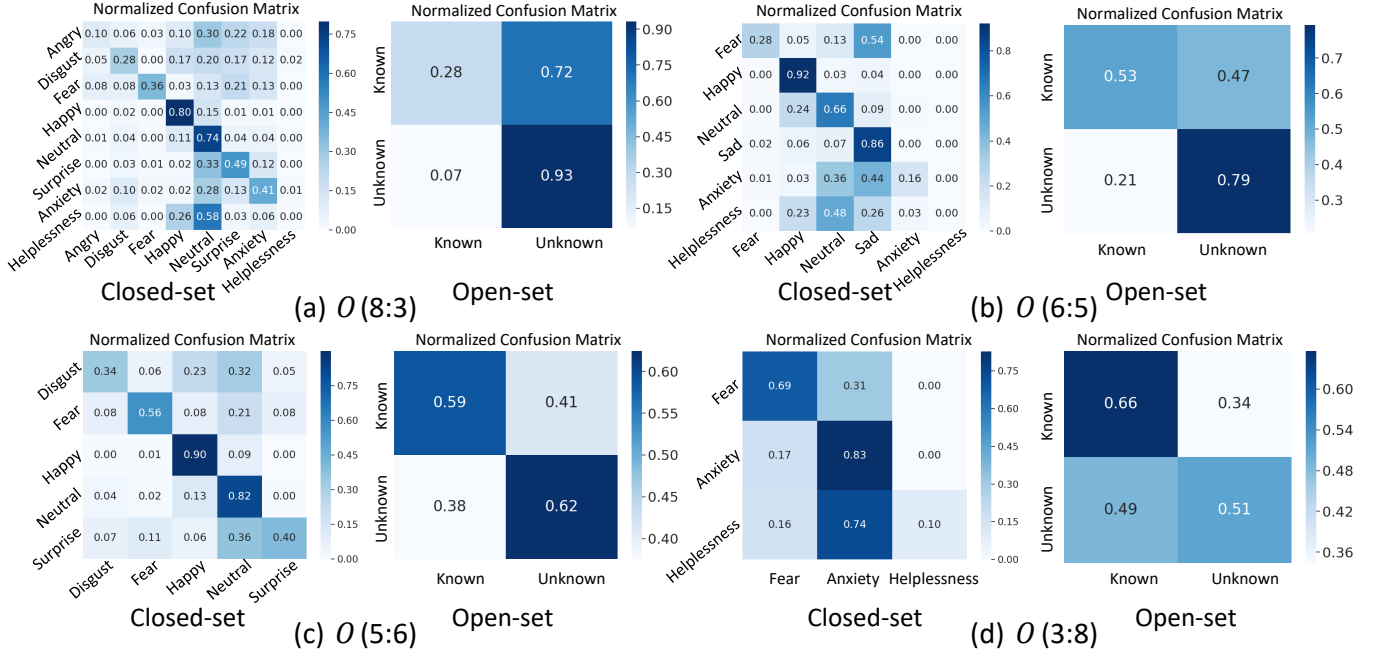


Figure 3: The confusion matrices of our method on the 11 emotion OV-FER task. Known and unknown classes are divided according to four different opennesses.

to learn more compact and discriminative features during training. Additionally, due to larger datasets and a greater number of emotion

categories in the other three tasks, current methods struggle to separate each category. In contrast, our method still distinguishes more feature categories than the others. This further demonstrates the

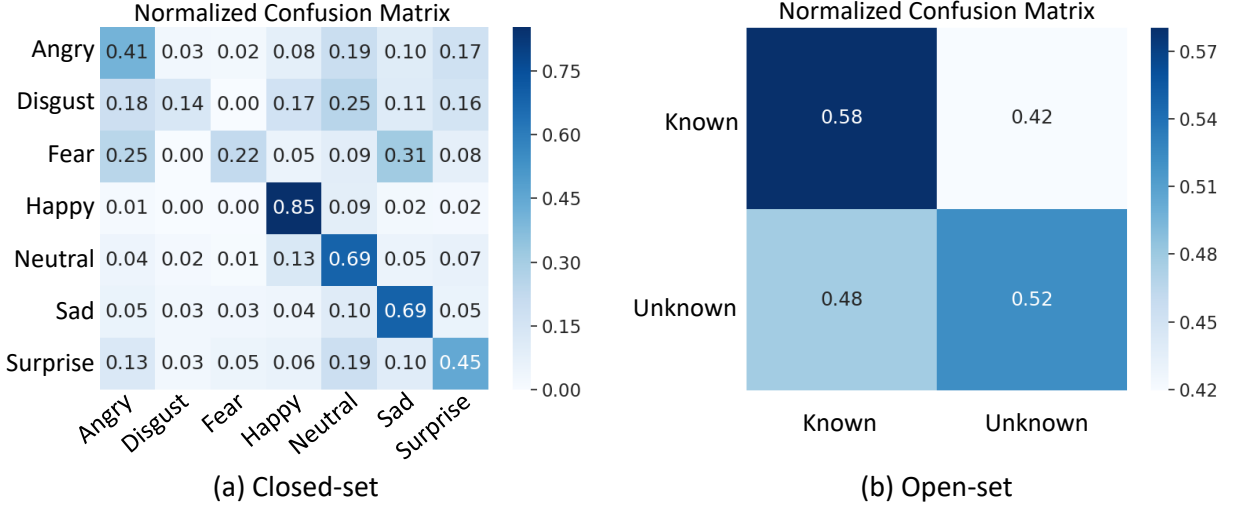


Figure 4: The confusion matrices of our method on the fusion dataset OV-FER task. 7 basic emotions in AFEW [2] and MAFW [3] are considered as known classes, while the other 4 emotions in MAFW are considered as unknown classes.

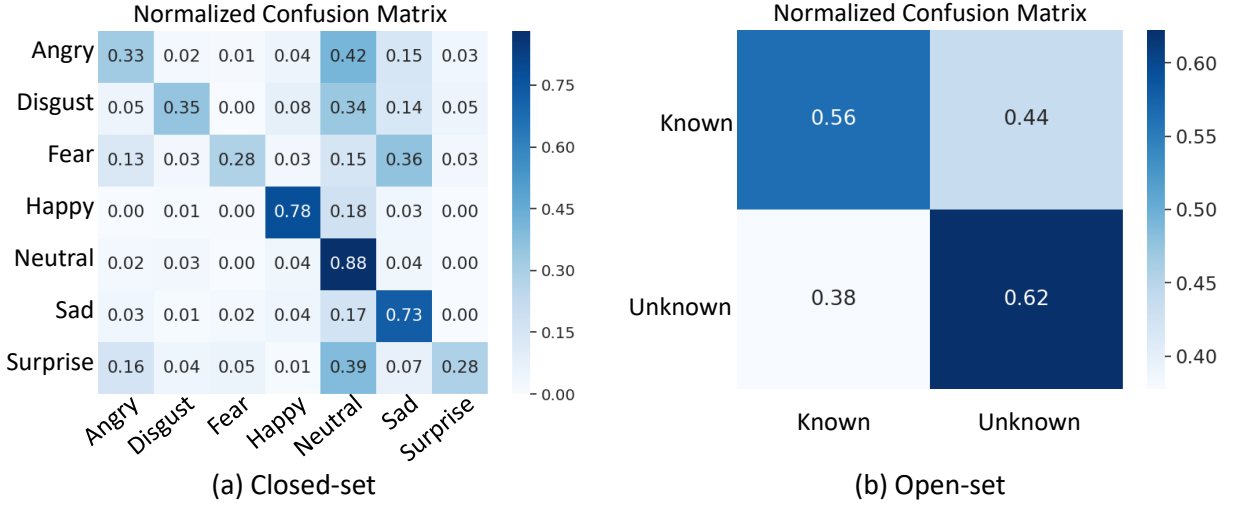


Figure 5: The confusion matrices of our method on the compound emotion OV-FER task. 7 basic emotions in MAFW [3] are used as known classes, and 9 compound emotions are used as unknown classes.

efficacy of our approach in extracting superior expression-sensitive features, effectively discerning between known and unknown categories, and accurately recognizing known categories.

2.3 Feature Maps and Prediction Results of Different Methods

To visually demonstrate the recognition ability of our method for known and unknown classes, Fig. 10 shows the results of predicting facial expressions using different methods, and visualizes the feature maps through CAM[7] with different methods. From the results, it is clear that our method not only correctly classified known classes, but also accurately identified unknown classes. Observing the feature attention maps, our method easily focuses on subtle

expression-sensitive areas, which proves that our proposed HESP effectively enhances CLIP and makes it well applied in OV-FER tasks.

References

- [1] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. 2021. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8065–8081.
- [2] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 423–426.
- [3] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*. 24–32.

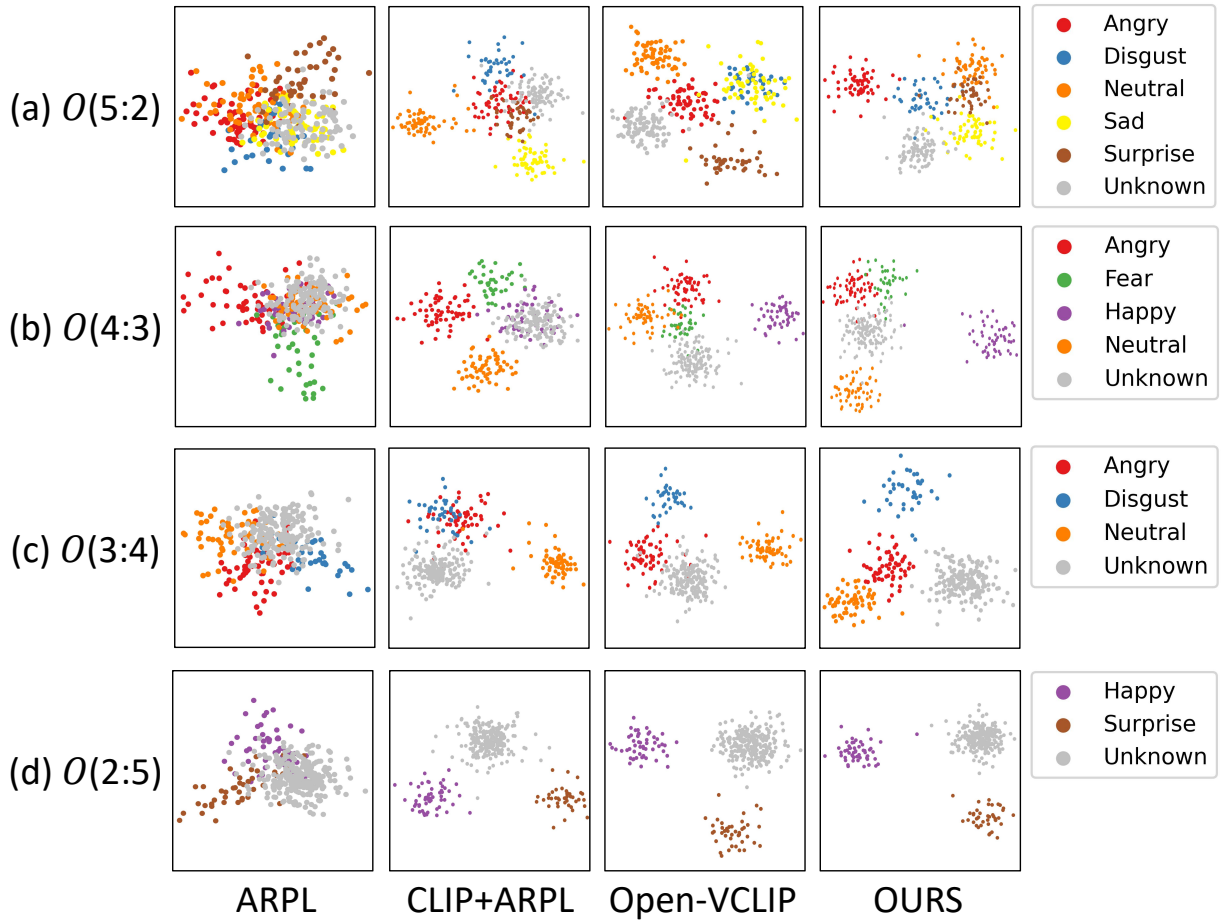


Figure 6: Visualization of facial expression features extracted by different methods on the 7 basic emotion OV-FER task.

- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [5] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6545–6554.
- [6] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. 2024. Building an open-vocabulary video CLIP model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [7] Bolei Zhou, Aditya Khosla, Agata Lapiedra, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

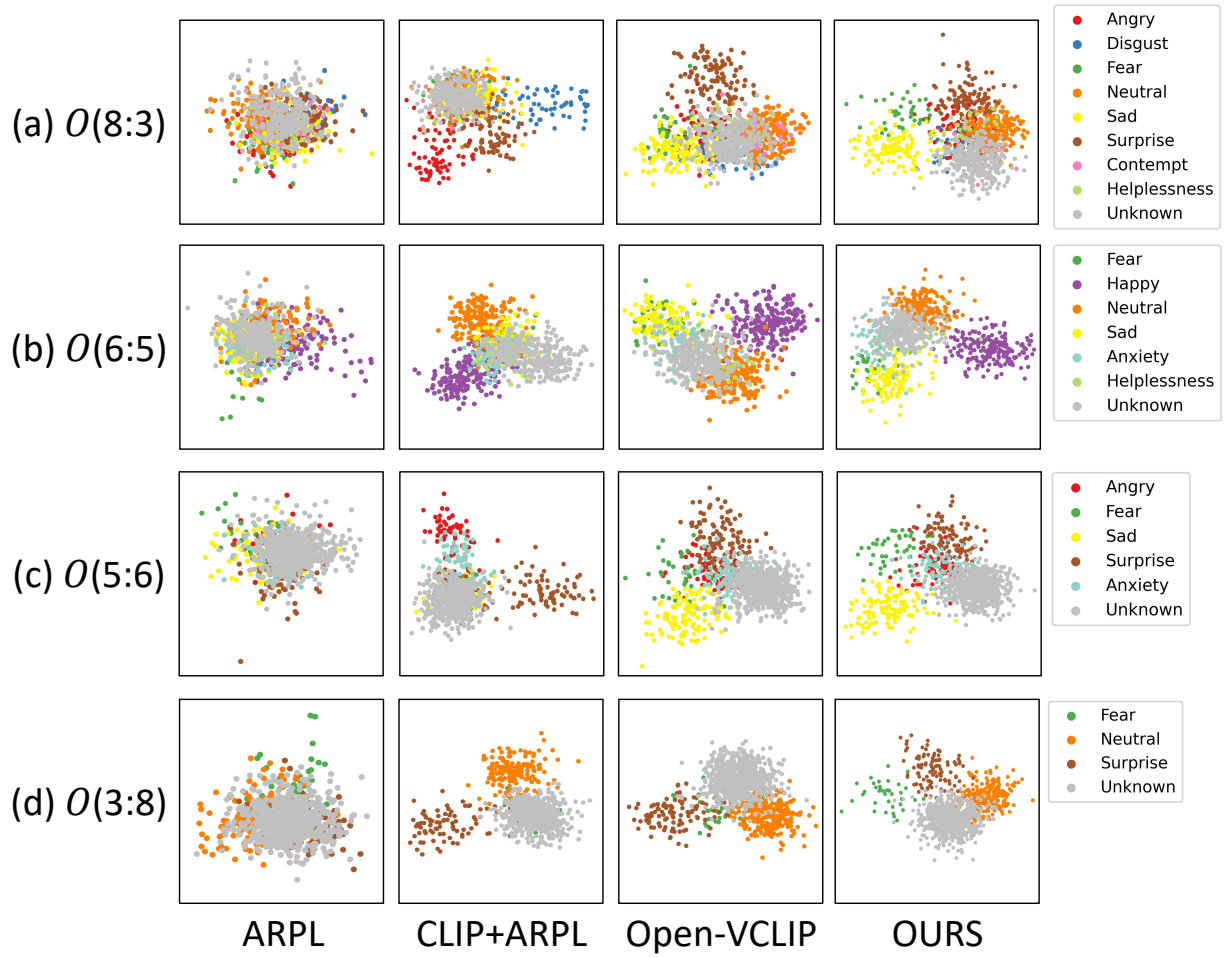


Figure 7: Visualization of facial expression features extracted by different methods on the 11 emotion OV-FER task.

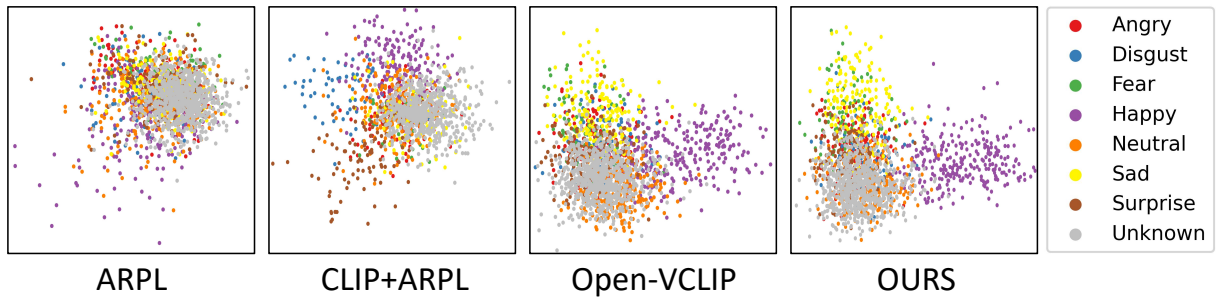


Figure 8: Visualization of facial expression features extracted by different methods on the fusion dataset OV-FER task.

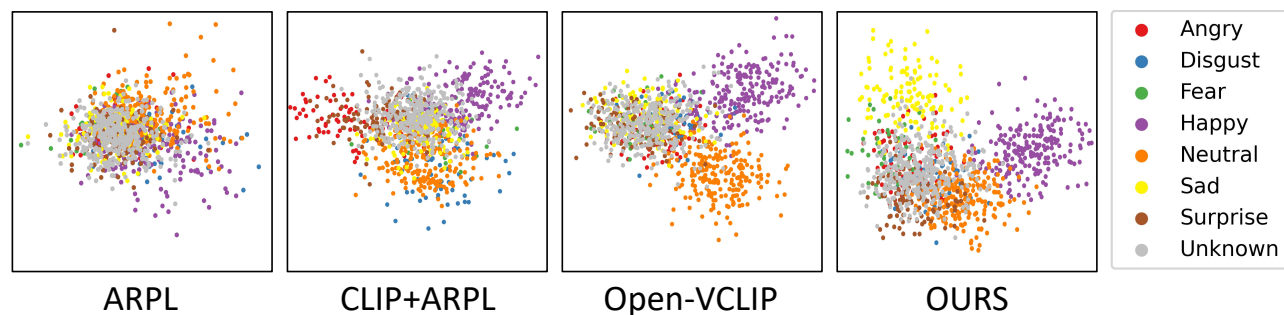


Figure 9: Visualization of facial expression features extracted by different methods on the compound emotion OV-FER task.

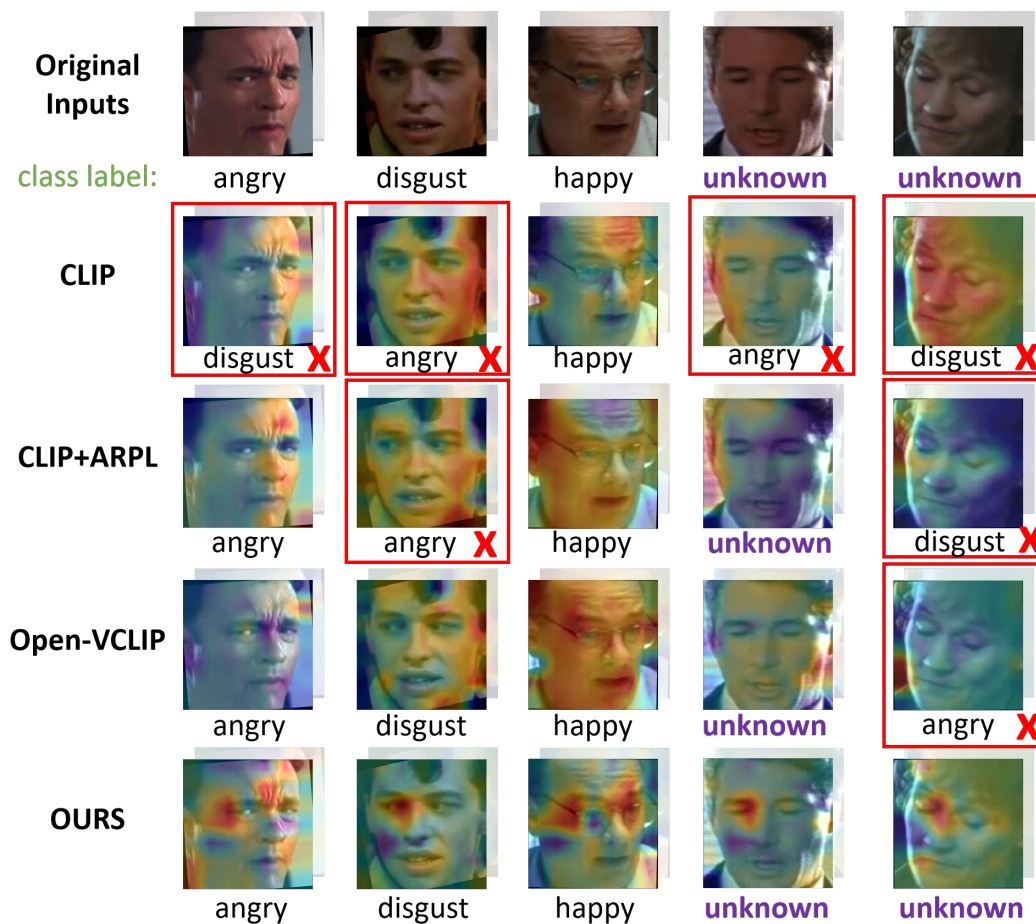


Figure 10: Prediction results and attended feature maps of different methods. Obviously, our method easily focuses on expression-sensitive areas, *e.g.*, eye areas, thus aiding in effectively discovering both known and unknown emotion information.