This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

GPT as Psychologist? Preliminary Evaluations for GPT-4V on Visual Affective Computing

Hao LU[1,2,*,†], Xuesong NIU[3,*,†], Jiyao WANG[1,2,*], Yin WANG[4,*], Qingyong HU[2,*], Jiaqi TANG[1,2,*], Yuting ZHANG[1,2], Kaishen YUAN[5], Bin HUANG[6], Zitong YU[5,‡], Dengbo HE[1,2], Shuiguang DENG[4], Hao CHEN[2], Yingcong CHEN[1,2,‡], Shiguang SHAN[7]
[1]The Hong Kong University of Science & Technology (Guangzhou),
[2]The Hong Kong University of Science & Technology, [3]Beijing Institute for General Artificial Intelligence,
[4]Zhejiang University, [5]Great Bay University, [6]Hangzhou Research Institute, Beihang University,
[7]Institute of Computing Technology, Chinese Academy of Sciences

Abstract

Multimodal large language models (MLLMs) are designed to process and integrate information from multiple sources, such as text, speech, images, and videos. Despite its success in language understanding, it is critical to evaluate the performance of downstream tasks for better human-centric applications. This paper assesses the application of MLLMs with 5 crucial abilities for affective computing, spanning from visual affective tasks and reasoning tasks. The results show that GPT-4V has high accuracy in facial action unit recognition and micro-expression detection while its general facial expression recognition performance is not accurate. We also highlight the challenges of achieving fine-grained micro-expression recognition and the potential for further study and demonstrate the versatility and potential of GPT-4V for handling advanced tasks in emotion recognition and related fields by integrating with task-related agents for more complex tasks, such as heart rate estimation through signal processing. In conclusion, this paper provides valuable insights into the potential applications and challenges of MLLMs in human-centric computing. Our interesting examples are at `https://github.com/EnVision-Research/GPT4Affectivity`.

1. Introduction

The development of multimodal large language models (MLLMs) has been a topic of growing interest in recent years [12, 22–24, 44]. MLLMs are designed to process and integrate information from multiple modalities, such as text, speech, images, and videos. The development of these models has been driven by the need to improve the accuracy and efficiency of various tasks, such as affective computing, sentiment analysis, and natural language understanding.

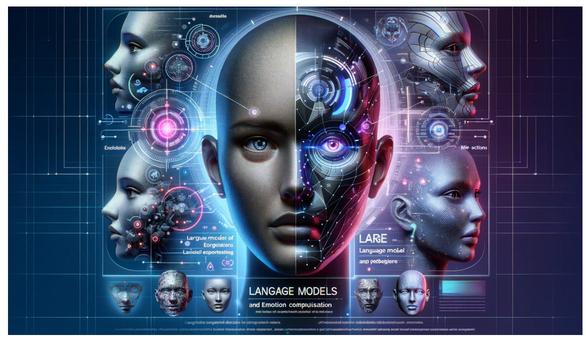

Figure 1. The propaganda image was generated by DALL·E 2.

MLLMs have shown great promise in improving the accuracy and robustness of affective computing systems [28, 38]. These models can process and integrate information from multiple modalities, such as facial expressions, speech patterns, and physiological signals, to infer emotional states accurately [12, 22, 24, 44] with significant implications for various applications, such as healthcare, education, and human-computer interaction.

Despite the rapid development of MLLMs, the need for standardized evaluation metrics is highlighted for accurate assessment. Different from the general language understanding evaluation benchmark that has been widely used to evaluate the performance of language models in NLP tasks [40–42], similar benchmarks to evaluate the performance of MLLMs for affective computing tasks are lacked, which is of great benefit to advance this field.

GPT-4V is the state-of-the-art MLLM that has shown remarkable success in various natural language processing tasks [1]. Its ability to process and integrate information

*Equal contribution.
†Project Leader.
‡Corresponding author.

from multiple modalities makes it an ideal candidate for evaluating the performance of MLLMs in tasks for affective computing. Furthermore, it can invoke a variety of tools that benefit affective computing tasks, such as related program generation with self-correction. For example, GPT-4V can call DALL·E 2 to generate high-quality visual affective images shown in Fig. 1.

In this paper, we evaluate GPT-4V with 5 typical human-centric tasks, spanning from visual affective tasks and reasoning tasks. We summarize our findings as follows:

(1) GPT-4V is highly accurate in recognizing facial action units. This accuracy can be attributed to its advanced understanding of facial movements and their corresponding emotions, which allows it to effectively identify and analyze facial action units.

(2) GPT-4V is also precise in detecting micro-expressions. Its ability to process subtle and transient facial expressions enables it to accurately capture these fleeting emotional cues, which are often difficult for humans to perceive.

(3) GPT-4V's performance in general facial expression recognition is not as accurate. This limitation may be due to the complexity and variety of facial expressions, resulting in the challenges in capturing and analyzing them. Nevertheless, when GPT-4V is used to process thought chains, its accuracy in facial expression recognition improves significantly. This improvement suggests that incorporating additional contextual information is of great importance to recognize facial expressions.

(4) Achieving high accuracy in micro-expression recognition remains a challenging task. This difficulty arises from the transient nature of micro-expressions and the need to detect and classify them within a very short time period. These challenges call for continuing research and development in this area for improving affective computing

(5) GPT-4V can also integrate with task-related agents to handle more complex tasks, such as detecting subtle facial changes and estimating heart rate with signal processing. By leveraging Python's powerful libraries and tools, GPT-4V can effectively process and analyze intricate facial data to derive valuable insights, such as heart rate estimation, which can further enhance its applications in mental health monitoring and virtual human companion systems.

2. Visual Affective Evaluation

Affective computing emerges as an interdisciplinary domain, leveraging computational technologies to discern, comprehend, and emulate human emotions. Its objective is to augment human-computer interaction, enhance user experiences, and facilitate improved communication and self-expression. Within the scope of computer vision, the analysis of human facial units [26], expressions [30], micro-expressions [48], micro-gestures [16], and deception detection [8], alongside physiological measurements [45], are pivotal to advancing emotional computing. Notably, large-scale pre-trained models, such as GPT-4V, have demonstrated substantial advancements in natural language processing, suggesting their considerable promise for application in affective computing. This study proposes to scrutinize the efficacy of GPT-4V across a variety of tasks, employing methodologies that include iterative conversations, open-ended inquiries, as well as multiple-choice and true/false questions.

2.1. Action Unit Detection

The Facial Action Coding System (FACS) [6] offers an explainable and reliable framework for the analysis of human facial expressions. It systematically deconstructs facial expressions into discrete components, known as Action Units (AUs), which correspond to the activation of specific facial muscles or groups thereof. Through the identification and quantification of these AUs, researchers can conduct a methodical examination of facial expressions and the emotional states they signify. Our assessment of GPT-4V's performance on the DISFA dataset [25], utilizing a gamut of question types, underscores its proficiency in accurately identifying AUs, thereby enabling precise emotion recognition from minimal interaction.

Remarkably, GPT-4V exhibits exceptional accuracy in AU identification, facilitating nearly flawless judgment across all AUs examined as shown in Tab. 1. Although our presentation includes a limited number of examples, our comprehensive evaluation reveals GPT-4V's surprising efficacy in this domain. To quantitatively appraise this performance, we adopted a quantitative analysis approach, benchmarking against the F1 metrics as reported in related studies.

Following [10], we report F1 metrics on DISFA. Specifically, we judge whether the recognition is successful by searching whether there is AUX (such as AU1) keyword in the reply question. The results show that the performance of GPT-4V is stronger than that of later professional models. This shows that GPT-4V has learned the micro-characteristics of emotion in a large number of network data and achieved significant recognition accuracy. Our findings indicate that GPT-4V's performance surpasses that of subsequent specialized models, underscoring its adeptness at learning the nuanced characteristics of emotion through extensive analysis of online data, thus achieving remarkable accuracy in emotion recognition.

2.2. Expression Recognition

The facial expression recognition [30] task involves identifying and analyzing human facial expressions to determine emotions. This task plays a crucial role in understanding human emotions, enhancing communication, and improv-

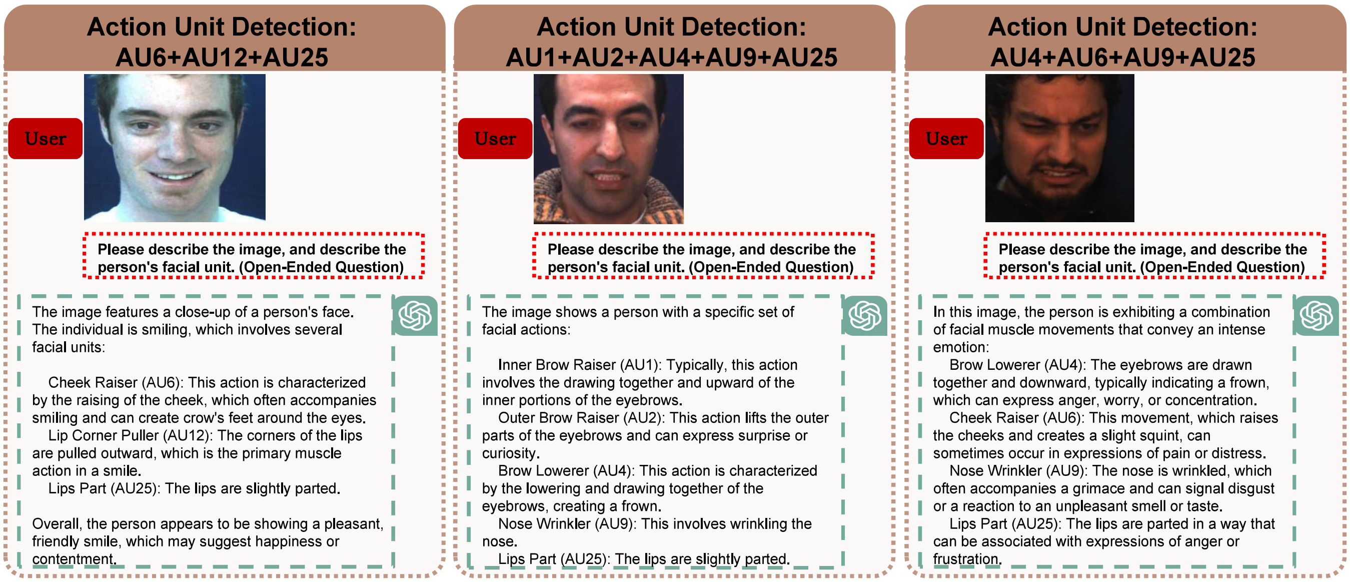

Figure 2. Action Unit detection on DISFA [25] dataset. We use the single round for the action unit. GPT-4V can accurately identify each AU.

AU	DRML [49]	DSIN [3]	LP [26]	SRERL [11]	EAC [13]	JAA [31]	ARL [32]	FAUDT [10]	PIAP [33]	ME-GraphAU [21]	BG-AU [4]	MPSCL [17]	GPT-4V
1	17.3	42.4	29.9	45.7	41.5	43.7	43.9	46.1	50.2	52.5	41.5	**62.0**	52.6
2	17.7	39.0	24.7	47.8	26.4	46.2	42.1	48.6	51.8	45.7	44.9	**65.7**	56.4
4	37.4	68.4	72.7	59.6	66.4	56.0	63.6	72.8	71.9	76.1	60.3	74.5	**82.9**
6	29.0	28.6	46.8	47.1	50.7	41.4	41.8	56.7	50.6	51.8	51.5	53.2	**64.3**
9	10.7	46.8	49.6	45.6	**80.5**	44.7	40.0	50.0	54.5	46.5	50.3	43.1	55.3
12	37.7	70.8	72.9	73.5	**89.3**	69.6	76.2	72.1	79.7	76.1	70.4	76.9	75.4
25	38.5	90.4	93.8	84.3	88.9	88.3	**95.2**	90.8	94.1	92.9	91.3	95.6	91.2
26	20.1	42.2	65.0	43.6	15.6	58.4	**66.8**	55.4	57.2	57.6	55.3	53.1	66.4
Avg.	26.7	53.6	56.9	55.9	48.5	56.0	58.7	61.5	63.8	62.4	58.2	65.5	**67.3**

Table 1. Comparison with state-of-the-art methods for AU detection on DISFA [25] dataset using the F1-score metric (in %).

ing mental health monitoring and virtual human companion systems. It can be challenging due to the complexity and variety of facial expressions, as well as the need to detect and classify subtle and transient expressions accurately. For this reason, we qualitatively analyze the performance of GPT-4V for emotion recognition on RAF-DB [30] dataset. Our methodology encompassed a multifaceted approach, employing iterative dialogues, open-ended questions, multiple-choice queries, and true/false assessments, specifically utilizing the CASME2 dataset as a basis for evaluation. Contrary to expectations, preliminary results indicate that GPT-4V exhibits limitations in accurately responding to even basic true/false questions related to emotion recognition, as depicted in the referenced figure shown in Fig. 3.

As shown in Fig. 3, natural emotions are thought to have no obvious characteristics, as soon as we pass a form of judgment question. For the emotion of Fear, GPT-4V thinks that the emotion is natural and cannot give the decision of fear. This is because emotions are inherently difficult to recognize without context, which is not considered an objective task. Therefore, GPT-4V cannot achieve good performance on these subjective tasks. This finding highlights a significant limitation in the application of advanced language models like GPT-4V for the nuanced task of emotion recognition. It suggests that while such models possess remarkable capabilities in various domains of natural language processing, their effectiveness in interpreting human emotions through facial expressions, especially in the absence of contextual information, remains constrained. The subjective nature of emotional expression, coupled with the subtleties and variations inherent in human facial expressions, necessitates a more sophisticated approach that incorporates contextual understanding and perhaps multimodal inputs that extend beyond textual analysis.

2.3. Compound Emotion Recognition

The task of compound emotion recognition [5] extends beyond the scope of simple emotion recognition by necessitating the identification and analysis of multiple emotions simultaneously exhibited through human facial expressions. The complexity of this task is amplified by the requirement to accurately detect and classify a spectrum of emotions, which may often be overlapping or present ambiguous signals. It can be more challenging than simple emotion recog-

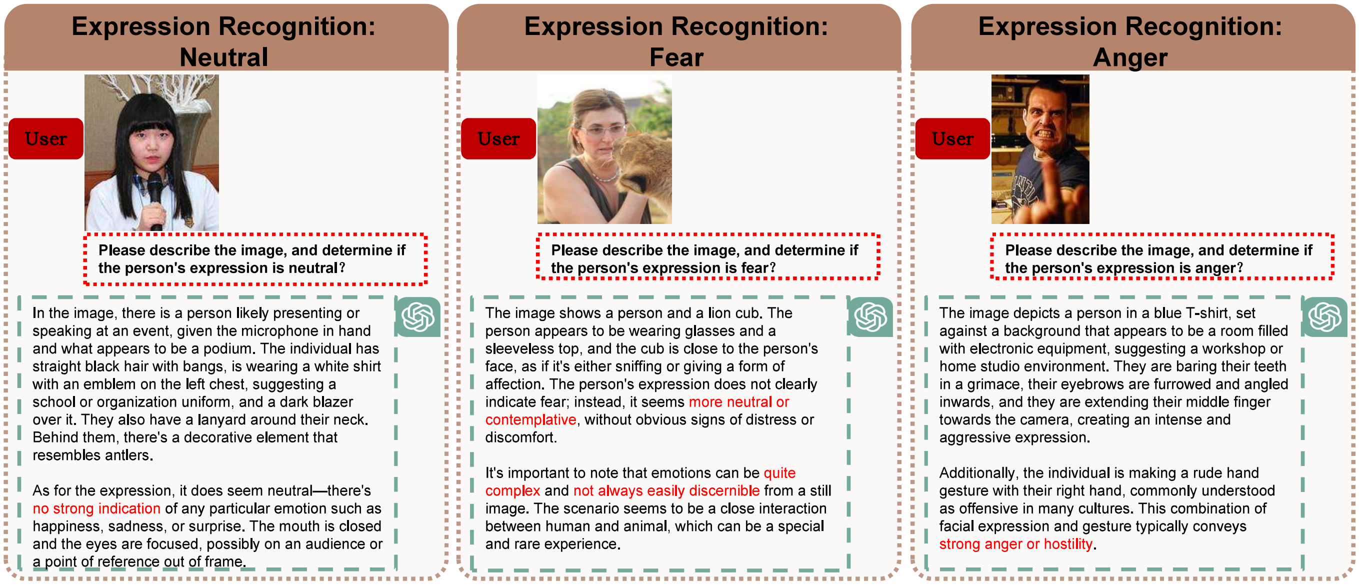

Figure 3. Expression recognition on RAF-DB [30] dataset. GPT-4V cannot achieve good performance on the subjective task of emotion recognition.

nition due to the need to detect and classify multiple emotions accurately, as well as the potential for conflicting or ambiguous expressions. In our continued exploration of GPT-4V 's capabilities, we extend our assessment to include the recognition of compound emotions.

As shown in Fig. 4, we qualitatively analyze the performance of GPT-4V for compound emotion recognition on RAF-DB [30] dataset and find that compound expressions can even be recognized. Even compound expressions are recognized more accurately than individual expressions. This does not mean GPT-4V is more accurate for compound than individual expressions. Instead, this is because the data of this compound expression is relatively more objective, and GPT-4V has an accurate judgment of this objective expression. This revelation underscores the importance of developing computational models that can navigate the intricacies of human emotions with a high degree of sensitivity and accuracy. For applications in mental health monitoring and virtual companionship, paving the way for innovations in emotional AI that can more closely mimic human empathetic and cognitive processes.

2.4. Micro-expression Recognition

The domain of micro-expression [48] research within emotion recognition is characterized by the endeavor to identify and interpret subtle, fleeting expressions that manifest on the human face. These micro-expressions, often resulting from rapid emotional shifts or attempts to conceal emotions, are particularly ephemeral, lasting only between 1/25 to 1/5 of a second. This attribute renders micro-expressions both a fascinating and formidable area of study [14, 15, 37]. However, the transient and elusive nature of micro-expressions presents significant challenges, notably in their detection and accurate interpretation. In our investigation, we meticulously crafted cue words and deployed a series of experimental setups involving judgment questions, multiple-choice inquiries, and iterative dialogues, all facilitated based on the CASME2 dataset [43] through the GPT-4V platform. This approach aimed to explore the potential of GPT-4V in recognizing and interpreting micro-expressions within the constraints of textual communication.

As shown in Fig. 5, GPT-4V did not answer the provided micro-expression test samples satisfactorily. GPT-4V cannot understand the difference between frames, and the difference is not visible to the human eye. We tried to amplify this difference, but GPT-4V thought the enlarged image was blurry, so GPT-4V was very weak on the microexpression task.

2.5. Micro-gesture Recognition

Micro-gesture recognition [16] tasks focus on recognizing and analyzing small, imperceptible body movements and facial expressions produced by people in specific scenarios, which usually represent an individual's inner emotions, attitudes, or cognitive responses. Micro-gesture recognition techniques are valuable for many applications, such as emotion recognition, negotiation, police interrogation, and mental health assessment. The core challenge of this technology is to capture brief and subtle changes in movements that are difficult for individuals to control due to their association with the autonomic nervous system. Micro-gesture recognition improves social interactions and communication by helping people better understand others' emotions

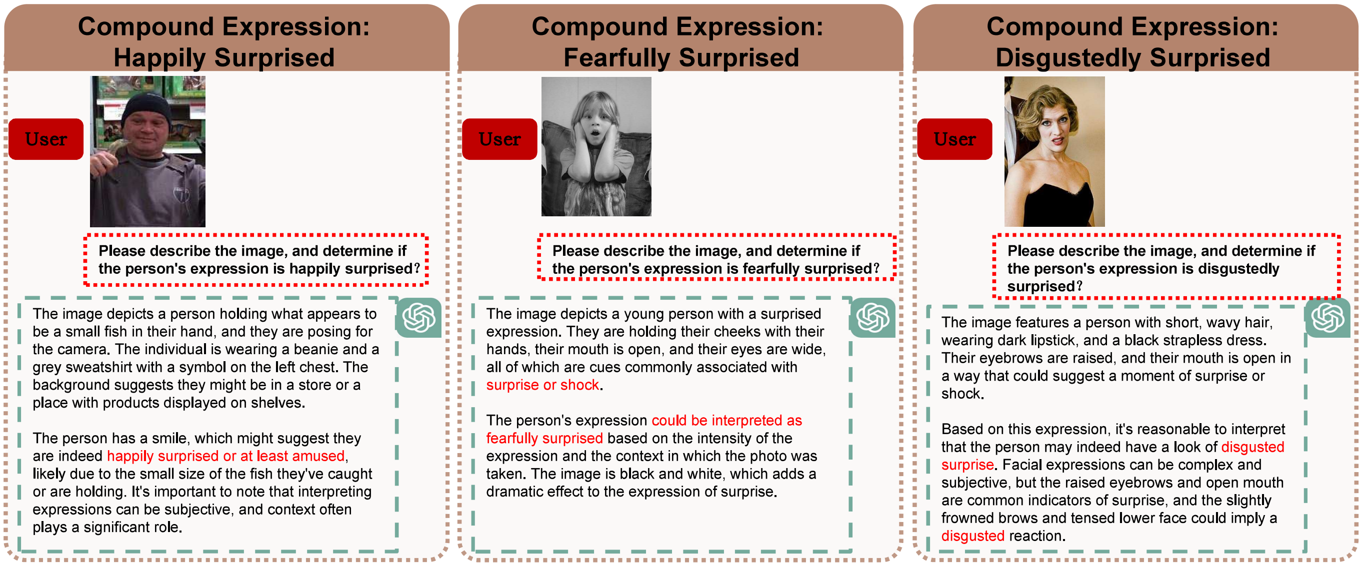

Figure 4. Compound emotion recognition on RAF-DB [30] dataset. GPT-4V can deduce objective compound expressions based on contextual information.

and motivations. We carefully designed the cue words and tested several different micro-gesture sequences on judgment questions, multiple-choice questions, and multi-round conversations using GPT-4V.

We qualitatively analyze the performance of GPT-4V for micro-gesture recognition on iMiGUE [16] dataset. As shown in Fig. 6, GPT-4V can give satisfactory answers to the micro-gesture test samples provided. It can give similar answers to even the most difficult questions (open-ended questions), such as rubbing the face (rubbing the eyes). As shown in Fig. 7, GPT-4V doesn't recognize the shoulder flutter. GPT-4V can't recognize tiny movements. While GPT-4V marks a significant step forward in the application of AI in the field of emotion and behavior recognition, its current limitations in recognizing certain micro-gestures suggest that further refinement and development is needed.

2.6. Deception Detection

Deception detection is an important task for determining the authenticity of video content, which is very important for security. To verify the performance of GPT-4V for deception detection, we evaluated on Real-Life Trial dataset [29].

As shown in 8, GPT-4V can't tell if a person in a video is lying. In fact, such subjective tasks are difficult for even real people to accurately judge. In addition, we try to input some multimodal information such as the sound spectrum to guide the GPT-4V to produce the correct result. But such operations do not allow GPT to reason the correct result. This shows that GPT-4V is still challenging for subjective tasks.

3. Advanced Capability of Reasoning

3.1. Chain of thought

The concept of Chain-of-Thought (CoT) [7, 34, 39] was first introduced in the seminal work by researchers at Google, titled Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. This innovative approach represents a significant advancement in cue strategies designed to enhance the performance of Large Language Models (LLMs) in executing complex reasoning tasks, encompassing arithmetic, common sense, and symbolic reasoning domains. Contrary to the Implicit Context Learning (ICL) approach, which primarily relies on input-output pairings, CoT incorporates a series of intermediate inference steps that scaffold toward the final solution, thereby enriching the model's reasoning pathway. In essence, CoT facilitates discrete prompt learning by appending an example to the beginning of a given input, enabling the model to process these concatenated texts simultaneously and produce the desired output. This method, under including additional intermediate prompts for inference, represents a substantial improvement over traditional context learning approaches.GPT-4V has a hard time recognizing specific expressions without context. However, we asked GPT-4V to first recognize the specific AU representation and then deduce the emotion based on the relationship between AU and the expression, which allowed GPT-4V to give some possible outcomes of the expression.

Furthermore, the application of CoT in emotion recognition tasks reveals its potential to circumvent some of the limitations faced by models such as GPT-4V in interpreting ambiguous or neutral expressions. Despite GPT-4V's pro-

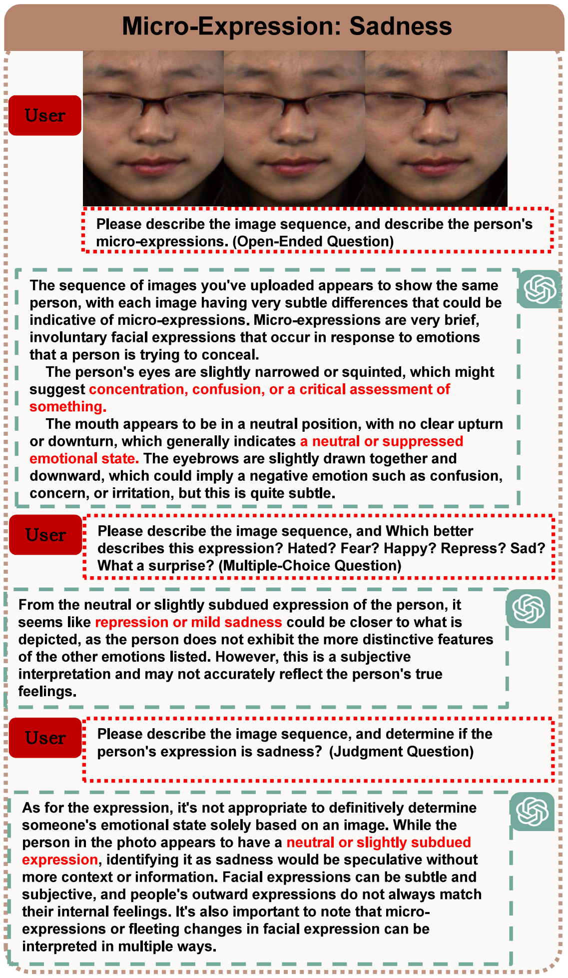

Figure 5. Micro-expression recognition on the CASME2 [43] dataset. GPT-4V has difficulty understanding the small differences in the image directly, so it is difficult to understand the micro facial expressions accurately.

ficiency in Action Unit (AU) recognition, its performance in emotion recognition from expressions remains suboptimal. By leveraging the correlation between CoT, AU, and facial expressions, we aim to enhance GPT-4V's accuracy in this area. As evidenced in Fig. 9, the incorporation of CoT significantly improves GPT-4V's capability to discern emotions, particularly in instances where expressions are ambiguous or lack clear contextual cues. This methodology enables GPT-4V to first accurately identify AUs, and subsequently infer the probable emotion based on the established relationship between AUs and facial expressions. The integration of CoT, as illustrated by the blue segments in the figure, thus facilitates a more nuanced understanding and recognition of emotional states by the model. Thus, the application of CoT in affective computing holds the potential to significantly improve the capability of visual

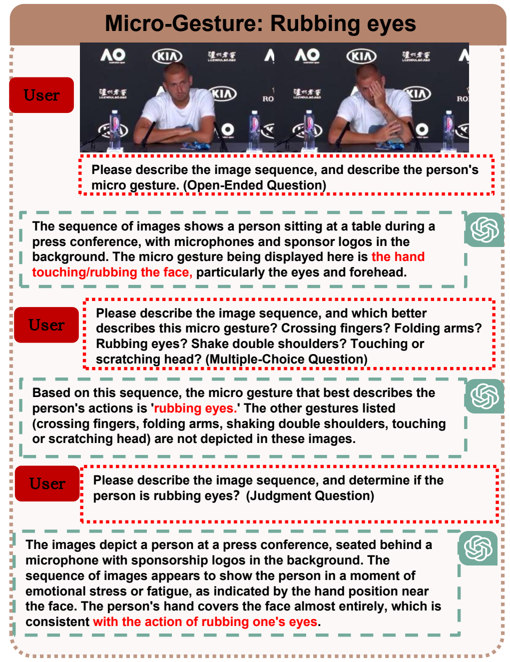

Figure 6. Micro-gesture (rubbing eyes) recognition on iMiGUE [16] dataset. GPT-4V can recognize obvious movements.

language models in interpreting and predicting emotional states with greater accuracy, leveraging contextual information to bridge the gap between task-related cues and the corresponding emotional expressions.

3.2. Tool call and processing

GPT-4V is one of the state-of-the-art multimodal language models that has achieved remarkable success in various natural language processing tasks. However, it is not directly applicable to some complex tasks, such as remote photoplethysmography (rPPG) [2, 9, 18–20, 27, 35, 36, 45–47]. rPPG is a non-invasive technique used to measure heart rate and respiratory rate from facial videos. It has a wide range of applications in healthcare, entertainment, and human-computer interaction.

Unfortunately, GPT-4V cannot read in long time-series videos and cannot discern subtle chromatic variations. To address this issue, a solution is for professional researchers to collaborate with GPT-4V. In this regard, we found that GPT-4V can call Python tools to run code and debug it. To demonstrate this process, we extracted facial video chromatic changes and used GPT-4V to process the signal. As shown in the figure, GPT-4V called Python to process and visualize the signal. During this process, there were several bugs, but GPT-4V was able to self-correct based on the bug information and ultimately provided an accurate heart rate result.

This process has provided us with an insight that this can be turned into a framework for a human-large language

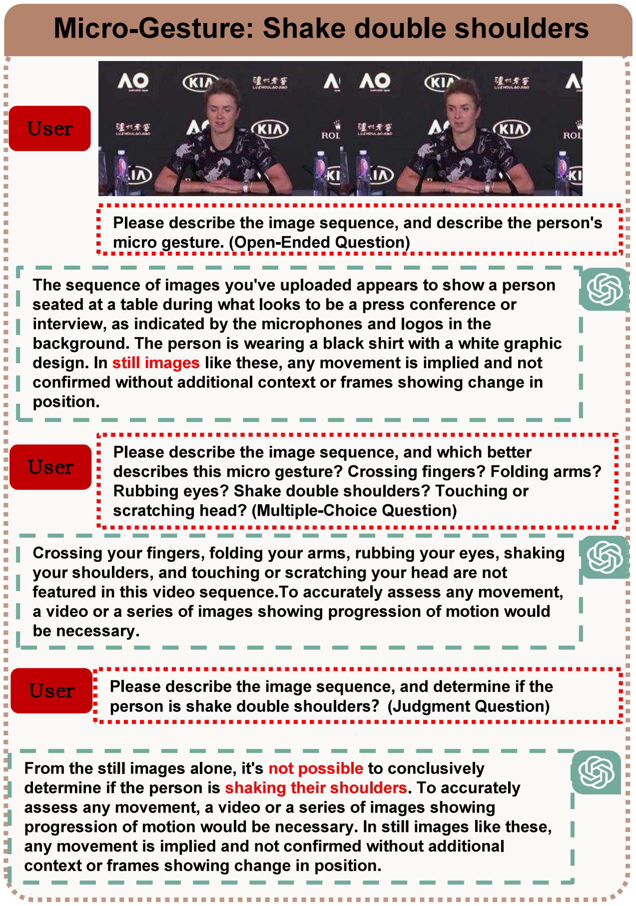

Figure 7. Micro-gesture (shaking double shoulders) recognition on iMiGUE [16] dataset. GPT-4V cannot recognize subtle movements.

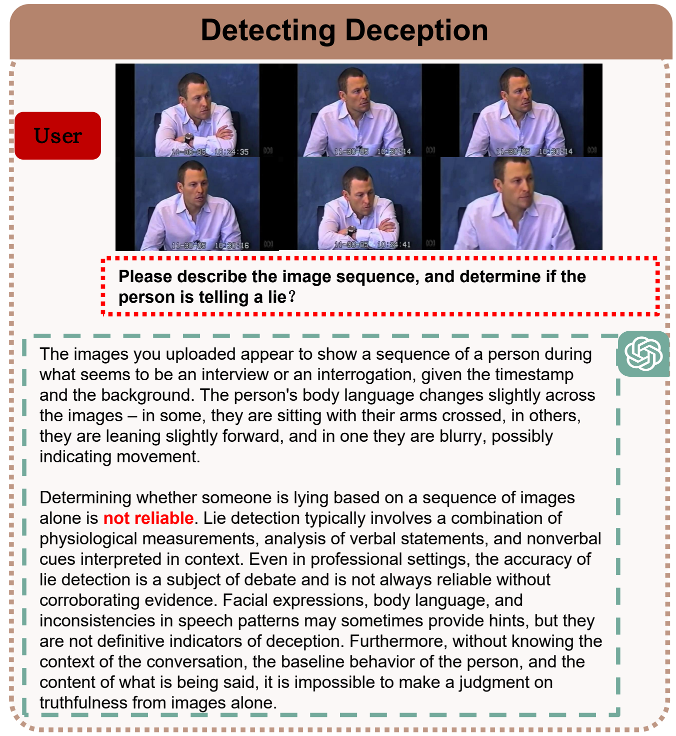

Figure 8. Deception Detection on Real-Life Trial dataset [29]. GPT-4V can't recognize a lie or not.

multimodal model that can self-correct. In this process, any large model can self-correct. This framework has immense potential in enhancing the accuracy and efficiency of various tasks, including rPPG.

4. Further Discussion

GPT-4V is a powerful language model that has shown remarkable success in various natural language processing tasks. However, it faces several challenges in other domains, such as facial expression recognition, emotion recognition, complex emotion recognition, non-contact physiological measurement, and authenticity detection.

Emotion recognition is the process of identifying and classifying emotions based on physiological signals, facial expressions, and speech patterns. GPT-4V has shown promising results in this task; however, it requires a large amount of training data and may not generalize well to new datasets. To overcome this limitation, future research can focus on developing transfer learning techniques that enable GPT-4V to learn from smaller datasets and generalize to new datasets.

Non-contact physiological measurement involves measuring physiological signals, such as heart rate, respiratory rate, and blood pressure, without direct contact with the body [2, 9, 19, 20, 35, 36]. GPT-4V faces difficulty in this task due to its limited ability to process and interpret physiological signals accurately. To overcome this limitation, future research can focus on developing new technologies that can capture physiological signals accurately and integrate them with GPT-4V.

Deception detection involves identifying and verifying the authenticity of a person or an object. GPT-4V faces difficulty in this task due to its limited ability to process and interpret visual and audio information accurately. To overcome this limitation, future research can explore ways to integrate GPT-4V with computer vision and audio processing techniques to improve authenticity detection accuracy.

In conclusion, GPT-4V faces several challenges in non-language tasks, such as facial expression recognition, emotion recognition, complex emotion recognition, non-contact physiological measurement, and deception detection. Future research can focus on developing new techniques to enhance GPT-4V's ability to process and integrate multimodal data and improve its accuracy and efficiency in these tasks.

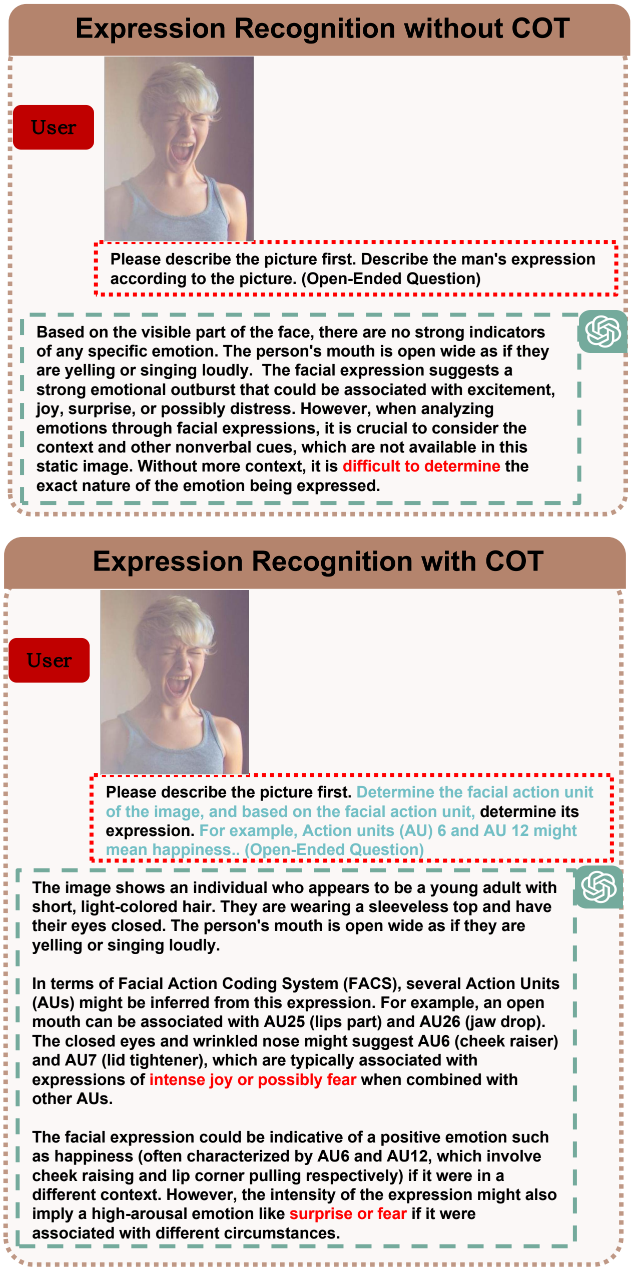

Figure 9. Improve the accuracy of expression recognition through COT. We use more accurate AU recognition and make GPT-4V recognition more accurate according to the relationship between AU and expression.

5. Conclusion

In this paper, we have discussed the challenges that GPT-4V faces in non-language tasks, such as facial expression recognition, emotion recognition, complex emotion recognition, non-contact physiological measurement, and authenticity detection. While GPT-4V has shown remarkable success in various natural language processing tasks, it faces limitations in these domains due to its limited ability to pro-

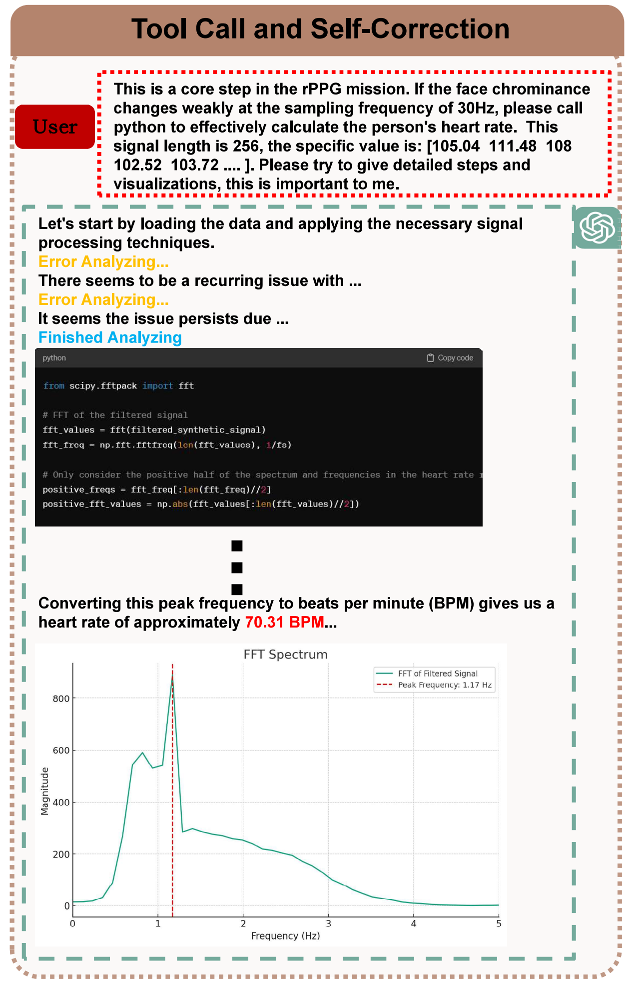

Figure 10. Tool call and processing for rPPG task. GPT-4V can directly write programs for signal processing on request and try to run them. If there is an error, GPT-4V can be further corrected according to the error prompt, and finally give the heart rate prediction and visualization.

cess and interpret visual and audio information accurately. To overcome these limitations, future research can focus on developing new techniques to enhance GPT-4V's ability to process and integrate multimodal data and improve its accuracy and efficiency in these tasks. This may involve integrating GPT-4V with computer vision and audio processing techniques, developing transfer learning techniques, exploring new sensor technologies, and improving the quality and quantity of training data. By addressing these challenges, GPT-4V has the potential to significantly advance the fields of facial expression recognition, emotion recognition, complex emotion recognition, non-contact physiological measurement, and authenticity detection, and open up new avenues for research and application in these domains. However, we should be careful about potential biases and privacy concerns, which will be a major consideration in the future.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 6, 7

[3] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the european conference on computer vision (ECCV)*, pages 298–313, 2018. 3

[4] Zijun Cui, Chenyi Kuang, Tian Gao, Kartik Talamadupula, and Qiang Ji. Biomechanics-guided facial action unit detection through force modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8694–8703, 2023. 3

[5] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014. 3

[6] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2

[7] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[8] Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Kot. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22135–22145, 2023. 2

[9] Bin Huang, Shen Hu, Zimeng Liu, Chun-Liang Lin, Junfeng Su, Changchen Zhao, Li Wang, and Wenjin Wang. Challenges and prospects of visual contactless physiological monitoring in clinical study. *NPJ Digital Medicine*, 6(1): 231, 2023. 6, 7

[10] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2, 3

[11] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8594–8601, 2019. 3

[12] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1

[13] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018. 3

[14] Yante Li, Xiaohua Huang, and Guoying Zhao. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing*, 436:221–231, 2021. 4

[15] Yante Li, Wei Peng, and Guoying Zhao. Micro-expression action unit detection with dual-view attentive similarity-preserving knowledge distillation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 4

[16] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10631–10642, 2021. 2, 4, 5, 6, 7

[17] Xin Liu, Kaishen Yuan, Xuesong Niu, Jingang Shi, Zitong Yu, Huanjing Yue, and Jingyu Yang. Multi-scale promoted self-adjusting correlation learning for facial action unit detection. *arXiv preprint arXiv:2308.07770*, 2023. 3

[18] Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia*, 2024. 6

[19] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12404–12413, 2021. 7

[20] Hao Lu, Zitong Yu, Xuesong Niu, and Ying-Cong Chen. Neuron structure modeling for generalizable remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18589–18599, 2023. 6, 7

[21] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *international joint conference on artificial intelligence*, 2022. 3

[22] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 1

[23] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

[24] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1

[25] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2, 3

[26] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang Shan. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11917–11926, 2019. 2, 3

[27] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019. 6

[28] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017. 1

[29] M Umut Şen, Veronica Perez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing*, 13(1):306–319, 2020. 5, 7

[30] Li Shan and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018. 2, 3, 4, 5

[31] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018. 3

[32] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 13(3):1274–1289, 2019. 3

[33] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 3

[34] Hao Wang, Euijoon Ahn, and Jinman Kim. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2431–2439, 2022. 5

[35] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, 2015. 6, 7

[36] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. 6, 7

[37] Yandan Wang, John See, Raphael C-W Phan, and Yee-Hui Oh. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12*, pages 525–537. Springer, 2015. 4

[38] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022. 1

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 5

[40] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023. 1

[41] Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*, 2020.

[42] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020. 1

[43] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014. 4, 6

[44] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1

[45] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021. 2, 6

[46] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022.

[47] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6):1307–1330, 2023. 6

[48] Guoying Zhao, Xiaobai Li, Yante Li, and Matti Pietikäinen. Facial micro-expressions: An overview. *Proceedings of the IEEE*, 2023. 2, 4

[49] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016. 3