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

Hao LU\textsuperscript{1,2,*,†}, Xuesong NIU\textsuperscript{3,*,†}, Jiyao WANG\textsuperscript{1,2,*}, Yin WANG\textsuperscript{4,*}, Qingyong HU\textsuperscript{2,*}, Jiaqi TANG\textsuperscript{1,2,*}, Yuting ZHANG\textsuperscript{1,2}, Kaishen YUAN\textsuperscript{5}, Bin HUANG\textsuperscript{6}, Zitong YU\textsuperscript{5,*}, Dengbo HE\textsuperscript{1,2}, Shuiguang DENG\textsuperscript{4}, Hao CHEN\textsuperscript{2}, Yingcong CHEN\textsuperscript{1,2,*}, Shiguang SHAN\textsuperscript{7}

\textsuperscript{1}The Hong Kong University of Science & Technology (Guangzhou), \textsuperscript{2}The Hong Kong University of Science & Technology, \textsuperscript{3}Beijing Institute for General Artificial Intelligence, \textsuperscript{4}Zhejiang University, \textsuperscript{5}Great Bay University, \textsuperscript{6}Hangzhou Research Institute, Beihang University, \textsuperscript{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 example is in 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.

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 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 AU X (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 improving 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 recognition due to the need to detect and classify multiple emo-
Expressions 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 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-
Micro-expression: Sadness

The sequence of images you've uploaded appears to show the same person, with each image having very subtle differences that could be indicative of micro-expressions. Micro-expressions are very brief, involuntary facial expressions that occur in response to emotions that a person is trying to conceal.

The mouth appears to be in a neutral position, with no clear upturn or downturn, which generally indicates a neutral or suppressed emotional state. The eyebrows are slightly drawn together and downward, which could imply a negative emotion such as confusion, concern, or irritation, but this is quite subtle.

As for the expression, it's not appropriate to definitively determine someone's emotional state solely based on an image. While the person in the photo appears to have a neutral or slightly subdued expression, identifying it as sadness would be speculative without more context or information. Facial expressions can be subtle and subjective, and people's outward expressions do not always match their internal feelings. It's also important to note that micro-expressions or fleeting changes in facial expression can be interpreted in multiple ways.

Micro-Gesture: Rubbing eyes

The images depict a person at a press conference, seated behind a microphone with sponsorship logos in the background. The sequence of images appears to show the person in a moment of emotional stress or fatigue, as indicated by the hand position near the face. The person's hand covers the face almost entirely, which is consistent with the action of rubbing one's eyes.

User

Please describe the image sequence, and which better describes this micro gesture? Crossing fingers? Folding arms? Rubbing eyes? Shake double shoulders? Touching or scratching head? (Multiple-Choice Question)

Based on this sequence, the micro gesture that best describes the person's actions is 'rubbing eyes.' The other gestures listed (crossing fingers, folding arms, shaking double shoulders, touching or scratching head) are not depicted in these images.

User

Please describe the image sequence, and determine if the person is rubbing eyes? (Judgment Question)

User

Please describe the image sequence, and determine if the person's expression is sadness? (Judgment Question)

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.

Figure 6. Micro-gesture (rubbing eyes) recognition on iMiGUE [16] dataset. GPT-4V can recognize obvious movements. However, 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...
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.
Expression Recognition without COT

User

Please describe the picture first. Describe the man’s expression according to the picture. (Open-Ended Question)

Based on the visible part of the face, there are no strong indicators of any specific emotion. The person’s mouth is open wide as if they are yelling or singing loudly. The facial expression suggests a strong emotional outburst that could be associated with excitement, joy, surprise, or possibly distress. However, when analyzing emotions through facial expressions, it is crucial to consider the context and other nonverbal cues, which are not available in this static image. Without more context, it is difficult to determine the exact nature of the emotion being expressed.

Expression Recognition with COT

User

Please describe the picture first. Determine the facial action unit of the image, and based on the facial action unit, determine its expression. For example, Action units (AUs) 6 and AU12 might mean happiness.

The image shows an individual who appears to be a young adult with short, light-colored hair. They are wearing a sleeveless top and have their eyes closed. The person’s mouth is open wide as if they are yelling or singing loudly.

In terms of Facial Action Coding System (FACS), several Action Units (AUs) might be inferred from this expression. For example, an open mouth can be associated with AU25 (lips part) and AU26 (jaw drop). The closed eyes and wrinkled nose might suggest AU6 (cheek raiser) and AU7 (lid tightener), which are typically associated with expressions of intense joy or possibly fear when combined with other AUs.

The facial expression could be indicative of a positive emotion such as happiness (often characterized by AU6 and AU12, which involve cheek raising and lip corner pulling respectively) if it were in a different context. However, the intensity of the expression might also imply a high-arousal emotion like surprise or fear if it were associated with different circumstances.

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 process 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.


[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


