Visual Prompting in LLMs for Enhancing Emotion Recognition

Qixuan Zhang[1*] **Zhifeng Wang**[1*] **Dylan Zhang**[2] **Wenjia Niu**[3]
Sabrina Caldwell[1] **Tom Gedeon**[1,4] **Yang Liu**[1†] **Zhenyue Qin**[5†]
[1]Australian National University [2]Quriosity Pty Ltd
[3]Webumate Pty Ltd [4]Curtin University [5]Yale University
{qixuan.zhang, zhifeng.wang, sabrina.caldwell}@anu.edu.au
info@quriosity.com.au, niuwenjia9064@hotmail.com
tom.gedeon@curtin.edu.au, {yang.liu1082, kf.zy.qin}@gmail.com
Project Page: https://wangzhifengharrison.github.io/sov_prompts_emnlp

Abstract

Vision Large Language Models (VLLMs) are transforming the intersection of computer vision and natural language processing. Nonetheless, the potential of using visual prompts for emotion recognition in these models remains largely unexplored and untapped. Traditional methods in VLLMs struggle with spatial localization and often discard valuable global context. To address this problem, we propose a Set-of-Vision prompting (SoV) approach that enhances zero-shot emotion recognition by using spatial information, such as bounding boxes and facial landmarks, to mark targets precisely. SoV improves accuracy in face count and emotion categorization while preserving the enriched image context. Through a battery of experimentation and analysis of recent commercial or open-source VLLMs, we evaluate the SoV model's ability to comprehend facial expressions in natural environments. Our findings demonstrate the effectiveness of integrating spatial visual prompts into VLLMs for improving emotion recognition performance.

1 Introduction

As the integration of computer vision and natural language processing progresses, VLLMs (Dai et al., 2024) are revolutionizing the way machines interpret visual and textual data. Emotion recognition is gaining considerable interest across multiple disciplines and presents distinct challenges (Yang et al., 2023a). It requires the decoding of emotions from nuanced indicators like facial expressions, body language, and contextual details.

Previous methods, such as those in (Xenos et al., 2024) and (Zhang et al., 2023b), enhance in-context emotion classification by training transformer-based models or CLIP to generate descriptions of emotions in visual contexts. However, these methods overlook the spatial relationships between different people and facial features within a single face. These spatial relationships can be labeled with numbers and bounding boxes in our SoV prompts to guide VLLMs. A person nearby may have similar facial expressions, while a person farther away may show different facial expressions. Additionally, the relationships between the eyes, mouth, and nose features can be highlighted by facial landmarks in SoV prompts to guide VLLMs.

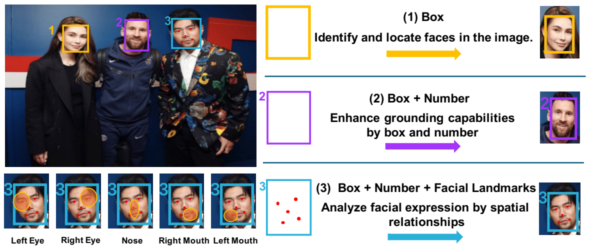


Figure 1: **Proposed Set-of-Vision (SoV) prompting approach for enhancing facial expression recognition in Vision-Language Large Models (VLLMs)**. SoV progressively incorporates (1) bounding boxes to identify and locate faces, (2) numbered boxes to ground and differentiate faces, and (3) facial landmarks to analyze spatial relationships for fine-grained emotion classification. This multi-stage visual prompting strategy enables VLLMs to accurately detect and recognize emotions in real-world images while preserving global context.

Recent studies (Yang et al., 2024; Zou et al., 2024; Yang et al., 2023b; Zou et al., 2024) have explored Visual Prompting, a technique employed in image-language tasks to guide LLMs by incorporating markers such as colorful boxes or circles to emphasize specific targets within an image. ReCLIP (Subramanian et al., 2022) adds colorful boxes directly onto an image to highlight specific targets and blurs other irrelevant areas to reduce the performance gap with supervised models on both real and synthetic datasets. Additionally, RedCircle (Shtedritski et al., 2023) employs visual prompt engineering, specifically drawing a red circle around an object in an image, to direct a Vision-Language

* Equal contribution
† Corresponding authors

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4484–4499
November 12-16, 2024

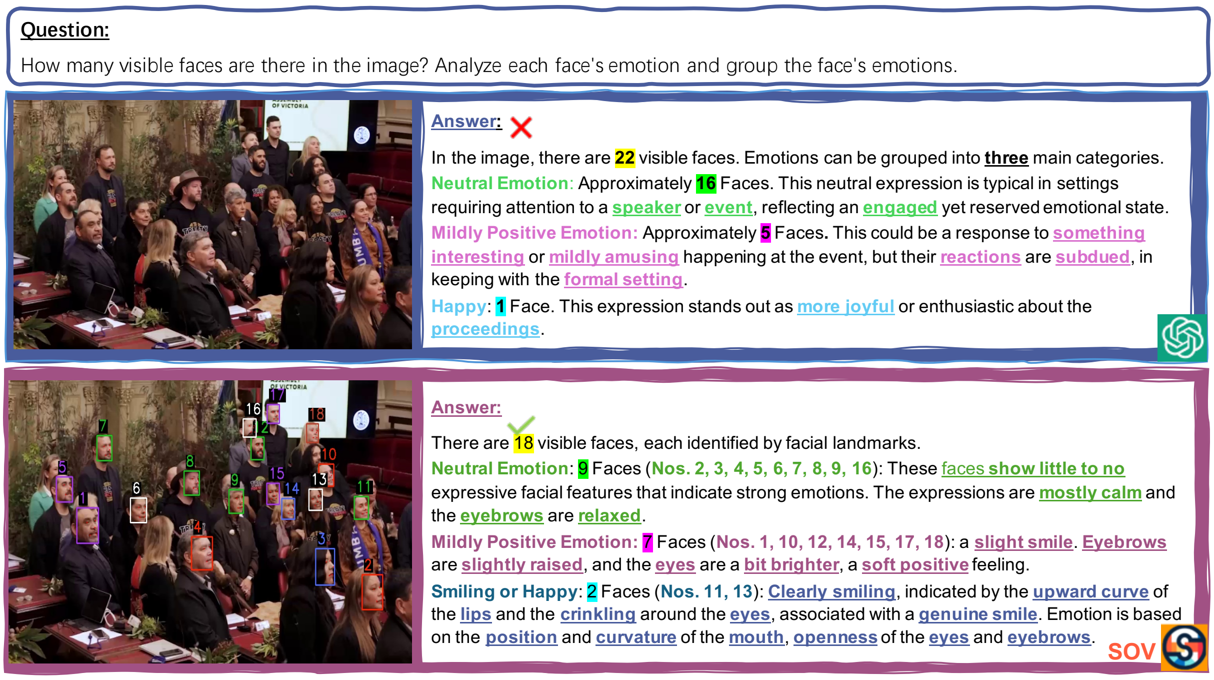


Figure 2: **Comparative analysis of emotion recognition methods in a group setting**: assessing the precision of facial emotion categorization and face detection using plain text prompts versus Set-of-Vision (SoV) prompts incorporating facial landmarks, bounding boxes, and face enumeration. **Top**: Results using plain text prompts. **Bottom**: Results using Set-of-Vision (SoV) prompts. The use of SoV prompts, such as numbering each face, placing bounding boxes, and identifying facial landmarks, allows for a more precise analysis.

Model's attention to that region and enhance its performance in tasks like zero-shot keypoint localization. However, both of these approaches focus on local objects and ignore spatial context information. Yang *et al.* (Yang et al., 2024) propose using fine-grained visual prompts, such as segmentation masks, and enhancing focus on relevant areas with a 'Blur Reverse Mask' that blurs regions outside the target mask to minimize distractions and maintain spatial context. Although visual prompting techniques have garnered interest, their full potential remains unexplored for emotion recognition tasks. Current approaches rely solely on coarse markers like colorful boxes, circles, or masks, which can introduce ambiguity, blur the face images, and pose challenges for accurate recognition tasks. This paper addresses these issues by systematically organizing and investigating various forms of visual prompting. Furthermore, we propose a new prompting approach called Set-of-Vision prompting (SoV) in Fig. 1, which utilizes spatial information such as numbers, bounding boxes, and facial landmarks to precisely mark each target while maintaining background context, thereby enhancing the zero-shot performance of facial expression recognition.

The top of Fig. 2, shows an approach where specific vision prompts are not used. As a result, the analysis inaccurately counts 22 visible faces and misclassifies persons' emotions into incorrect categories, with a number of faces labeled under 'Neutral Emotion' and fewer under 'Mildly Positive Emotion' and 'Happy'. This misclassification and miscount demonstrate the limitations when detailed visual cues are not utilized in the analysis. In the bottom of Fig. 2, the use of SoV prompts, such as numbering each face, placing bounding boxes, and identifying facial landmarks, allows for a more precise analysis. The correct number of faces is identified (18), and the emotions are accurately categorized into more nuanced groups: 'Neutral Emotion', 'Mildly Positive Emotion', and 'Smiling or Happy'. This method provides a clearer and more detailed breakdown of each individual's emotional state based on visible facial expressions. This comparison highlights the importance and effectiveness of integrating visual prompts in VLLMs analysis for more accurate and detailed recognition and categorization of human emotions in images.

To summarize, our main contributions are: (1) The paper introduces a novel visual prompting method (SoV) that highlights facial regions directly within the entire image. This preserves background context, enhancing the ability of VLLMs to perform accurate emotion recognition without

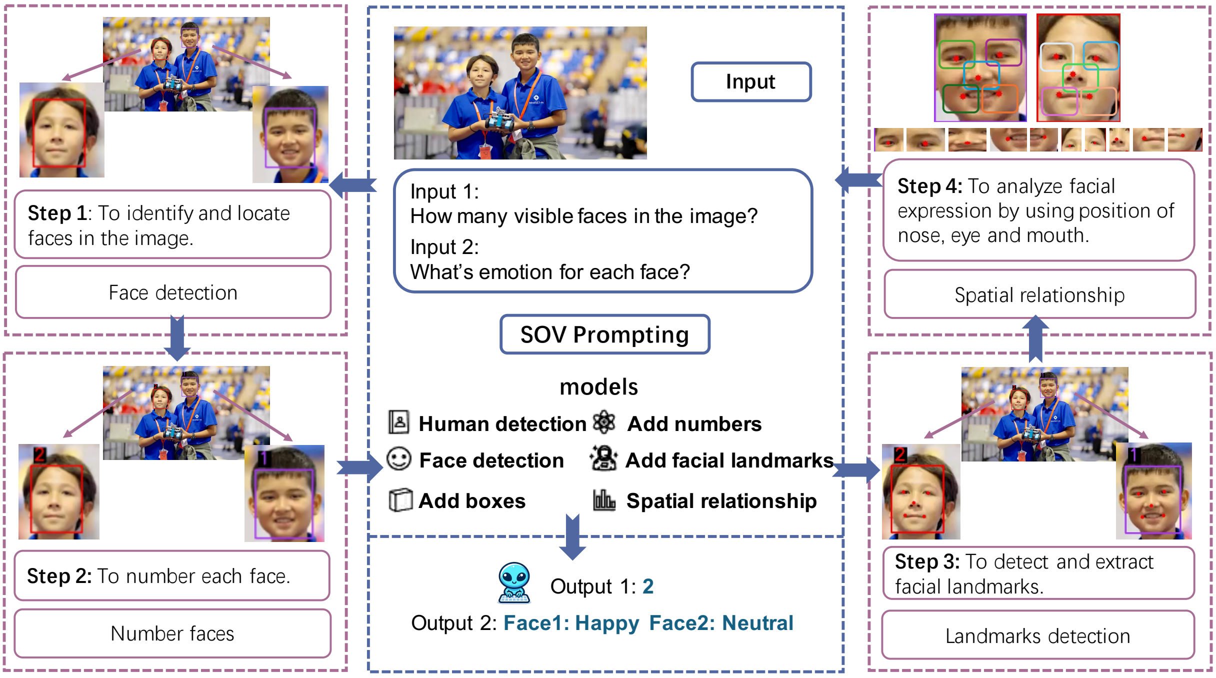


Figure 3: **Workflow diagram for enhanced face recognition and emotion analysis using the Set-of-Vision (SoV) prompting approach**: a multi-step process involving face detection, face numbering, landmark extraction, and spatial relationship analysis for emotion classification. Each detected face is analyzed and identified by facial landmarks on the face, such as the positions of the nose, eyes, mouth, and other facial features.

the need for cropping faces, thus maintaining the holistic view of the image. (2) The proposed face overlap handling algorithm effectively addresses conflicts arising from overlapping face detections, especially in images with dense face clusters. By prioritizing larger faces and iteratively checking for overlaps, the algorithm ensures that non-occluded faces are retained for subsequent emotion analysis. (3) Our results show that incorporating spatial visual prompts (SoVs) into VLLMs can enhance their performance in recognizing emotions.

2 Related Work

2.1 Vision Large Language models

LLMs such as LLaMA (Touvron et al., 2023), ChatGPT-3 (Brown et al., 2020), ChatGPT-4 (Achiam et al., 2023), and PaLM (Chowdhery et al., 2023) have demonstrated remarkable zero-shot transfer capabilities in natural language processing. Recently, VLLMs, which leverage image-text data pairs from the web, have gained prominence in the computer vision domain. MiniGPT-4 (Zhu et al., 2023), a model that combines a visual encoder with an advanced language model, can enable multi-modal capabilities such as generating detailed image descriptions and designing websites from sketches. Video-LLaVA (Zhang et al., 2023a) is a multi-modal framework that enhances Large Language Models with the ability to understand and generate responses based on both visual and auditory content in videos. LLaVA (Liu et al., 2023) is a newly developed, end-to-end trained, large multimodal model that combines a vision encoder with a language model, demonstrating promising abilities in multimodal chat. Although VLLMs exhibit remarkable capabilities in vision-based tasks such as image segmentation and object detection, they typically require fine-tuning of the vision and text encoders using existing open vocabulary methods when applied to specific tasks. In contrast, this paper proposes a zero-shot architecture for emotion recognition, overcoming the need for task-specific fine-tuning.

2.2 Prompting methods

Prompt engineering is a widely employed technique in the field of NLP (Strobelt et al., 2022; Zhou et al., 2024). AdbGPT (Feng and Chen, 2024) is a novel, lightweight approach that leverages few-shot learning and chain-of-thought reasoning in Large Language Models to automatically reproduce bugs from bug reports, mimicking a developer's problem-solving process without the need for training or hard-coding. Although prompts for large language models have been extensively explored, prompts for vision tasks have received less attention and investigation. Yang *et al.* (Yang et al., 2024) propose using fine-grained visual prompts like segmentation masks and a Blur Reverse Mask strategy to focus on relevant areas. The Image-of-Thought (IoT) prompting method (Zhou et al., 2024) enhances Multimodal Large Language Models by guiding them to extract and refine visual rationales step-by-step from images, combining visual and textual insights to improve zero-shot performance on complex visual reasoning tasks. Although these methods have shown promise in tasks like semantic segmentation and object grounding, their performance in emotion recognition has been less effective. This is largely because they tend to analyze individual objects in isolation, overlooking global information and specific facial features, which are crucial for accurately interpreting emotions. To address these issues, the proposed approach directly focuses on the fine-grained facial features present in the entire image, preserving spatial information by utilizing bounding boxes, numbers, and facial landmarks.

3 Methods

3.1 Problem Definition

The task of matching images to emotions for each visible face in a given image involves several sophisticated steps, combining face detection and emotion recognition. Typically, the VLLMs, denoted as Φ, will take an image $I \in \mathcal{R}^{H\times W\times 3}$ and a text question of length l_i, $Q^i = [q^i_1, q^i_2, ..., q^i_{l_i}]$, as input. The output is a sequence of answers with length l_o, containing emotions, $A^o = [a^o_1, a^o_2, ..., a^o_{l_o}]$, which can be formulated as (Eq. 1):

$$A^o = \Phi(I, Q^i) \tag{1}$$

In our task, we aim to find the best matching image-emotion pairs (I, A^o) for each visible face. Traditionally, this involves cropping the face from the image using face detectors. However, with the introduction of visual prompting, faces can be directly marked on the entire image, highlighting the facial region while preserving the background context and avoiding the obscuration of faces. With this in mind, we have developed Set-of-Vision prompts (SoV), a simple method of overlaying a number of visual prompts on the facial regions in an image.

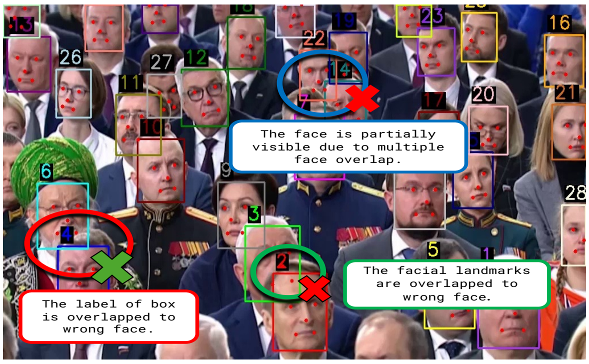


Figure 4: **Face detection inevitably introduces some overlaps or conflicts that confuse VLLMs.** Analyzing the impact of face overlaps, occlusions, landmark misalignment, and bounding box conflicts for emotion recognition.

Algorithm 1 Detect and Handle Overlaps

// Define a function to check if overlap
def Boxes_overlap(B_1, B_2):
  $\tilde{O} = Check\_Overlap(B_1, B_2)$ // Check if overlap
  return $\tilde{O}$
// Define a function to calculate the area
def Face_size(B):
  $\tilde{A} = Calculate\_Area(B)$ // Calculate the area
  return $\tilde{A}$
// Main function
def Detect_and_handle_overlaps(faces):
  // Sort faces by the area in descending order
  $\tilde{F} =$ **Sorted**$(faces)$
  // Initialize an empty list
  visible_faces = []
  for k in range(K)
    if Boxes_overlap($\tilde{F}[k], \tilde{F}[k-1]$):
      if Face_size($\tilde{F}[k]$) $>$ Face_size($\tilde{F}[k-1]$):
        // Replace the smaller face
        visible_faces $= \neg\tilde{F}[k-1] \wedge \tilde{F}[k]$
  return visible_faces

This operation augments the input image I to a new image $I^{new} = SoV(I)$, while keeping the text prompts to VLLMs unchanged as shown in Fig. 2. It can be formulated as (Eq. 2):

$$A^o = \Phi(SoV(I), Q^i) \tag{2}$$

3.2 Set of Vision Prompts

3.2.1 Box detection

Once the image is obtained, we need to generate visual prompts for the image that will be utilized by VLLMs for emotion recognition. We employ the RetinaFace (Deng et al., 2020) algorithm to detect faces within the image. Let $B = \{b_1, b_2, \ldots, b_n\}$ denote the set of detected face bounding boxes, where b_i represents the i-th face bounding box. The process can be formulated as (Eq. (3)):

$$b_i = \mathcal{D}(I, \theta_i) \tag{3}$$

, where I is the input image; θ_i represents the hyperparameters for the RetinaFace model $\mathcal{D}$; and b_i corresponds to the i-th face bounding box.

3.2.2 Box Overlap Handling Algorithm

However, this face detection algorithm inevitably introduces some overlaps or conflicts that confuse VLLMs, especially in images with densely populated faces, such as when two faces overlap in one area or one face is obscured by another. This is illustrated in Fig. 4. To mitigate this problem, we propose a face overlap handling algorithm, as shown in Algorithm. 1. Given the set of boxes $B = \{b_1, b_2, \ldots, b_n\}$, we first calculate the area for each bounding box b_i, then sort the detected faces b_i by their area in descending order (line 12) (Eq. (4)):

$$B_{sorted} = \{b_1, b_2, \ldots, b_n\} \tag{4}$$

ensuring $\text{Area}(b_1) \geq \text{Area}(b_2) \geq \ldots \geq \text{Area}(b_n)$. It ensures that larger faces are prioritized.

By iterating through the sorted faces, the algorithm checks for overlaps and compares the areas of overlapping faces. For each face $\tilde{F}[k]$ in B_{sorted}, it checks if $\tilde{F}[k]$ overlaps with any face $\tilde{F}[j]$ in F_{final} by (Eq. (5)):

$$\begin{aligned}&\text{Overlap}(\tilde{F}[k], \tilde{F}[j]) = \\ &\frac{\text{Area}(\tilde{F}[k] \cap \tilde{F}[j])}{\min(\text{Area}(\tilde{F}[k]), \text{Area}(\tilde{F}[j]))} > \epsilon\end{aligned} \tag{5}$$

If the overlap is significant, compare their areas and discard the smaller face by (Eq. (6)):

$$\tilde{F}[k] = \begin{cases} \hat{F}[k] & \text{if Area}(\hat{F}[k]) > \text{Area}(\hat{F}[j]), \\ \hat{F}[j] & \text{otherwise.} \end{cases} \tag{6}$$

Add non-occluded faces to F_{final} by (Eq. (7)):

$$F_{final} \leftarrow F_{final} \cup \{\tilde{F}[k]\} \tag{7}$$

It ensures that only the faces that are close to the camera and not obstructed remain in the final list.

Once we determine the location of boxes $B = \{b_1, b_2, \ldots, b_n\}$ for each face, We need to assign a unique ID to each face. The unique ID $N = \{1, 2, \ldots, n\}$ will be used to locate each face in the image, where n is the number of detected faces. Thus, the set of vision prompts becomes distinguishable and can be effectively interpreted by VLLMs.

3.2.3 Facial Landmarks Detection and Analysis

After identifying and handling overlapping faces, we proceed with the detection and extraction of facial landmarks for each face. For each face $b_i \in F_{final}$:

$$\begin{aligned}L_i &= \Theta(b_i) \\ &= \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}\end{aligned} \tag{8}$$

, where Θ is landmarks extraction model, L_i is the facial landmarks, (x_m, y_m) is coordinates of facial landmarks. For each set of landmarks L_i, analyze spatial relationships for facial expression within RGB image:

$$E(L_i) = f(R(L_i)) \tag{9}$$

, where f is a function mapping spatial relationships $R(L_i)$ to facial expressions $E(L_i)$. The entire process is illustrated in Fig. 3.

3.3 Text and Vision Prompts

We have a collection of n pairs of location-vision prompts, represented as $(l_1, v_1), \ldots, (l_n, v_n)$. When introducing additional text prompts for a new image I_{new}, we can choose to use either plain text prompts or a combination of text and vision prompts.

3.3.1 Plain Text Prompts

This method is exemplified on the left side in Fig. 5. It involves asking a general question about the emotional state of a group of people without referencing specific individuals. For example, the question "What is the emotion for this group of people?" yields an answer that considers the overall mood and setting of the group. This approach is useful for understanding group dynamics or the general atmosphere of a scene.

3.3.2 Combined Text-Vision Prompts

Shown on the right side in Fig. 5, this method involves more detailed prompts that focus on individual persons within the group. This allows for a more nuanced analysis of specific people's emotions and actions. For instance, questions such as 'What is Person 1's emotion?' or 'What food is Person 3 eating and drinking?' prompt answers that delve into specific details regarding individuals' facial expressions, body language, and interactions with objects, like food and drinks.

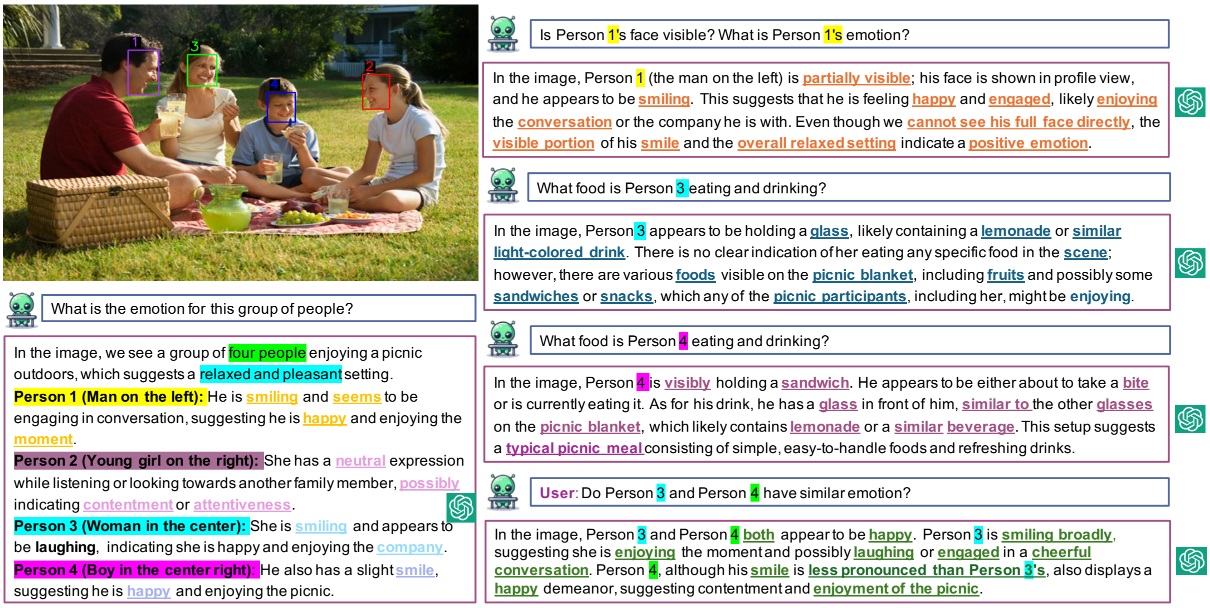


Figure 5: We use two types of prompt methods. **Left**: plain text prompts, which can be used for group emotion recognition. **Right**: combined text-vision prompts, which can be used for analyzing specific individuals' emotions. These prompts can be used to evaluate emotional interpretation in social interactions based on facial expressions, body language, and contextual cues.

Table 1: **Comparison of zero-shot emotion recognition methods**, including MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023), Video-LLaVA (Zhang et al., 2023a), GPT-4V (Achiam et al., 2023), and SoV-Enhanced GPT Models, across datasets with varying difficulty levels (Easy, Medium, and Hard): A Comparative Analysis of Accuracy and Top-1 Recall (R@1).

Methods	Backbone	Easy		Medium		Hard		Total	
		Acc (%)	R@1	Acc (%)	R@1	Acc (%)	R@1	Acc (%)	R@1
MiniGPT-4 (Zhu et al., 2023)	Q-former, ViT	30.45	16.17	19.88	12.85	15.78	14.10	22.87	12.96
LLaVA (Liu et al., 2023)	CLIP, ViT	35.74	15.91	22.80	11.29	3.50	1.58	22.65	10.56
Video-LLaVA (Zhang et al., 2023a)	Pre-align ViT	20.11	9.37	16.95	7.26	8.77	4.46	16.12	6.84
GPT-4V (Achiam et al., 2023)	ViT	48.85	27.94	47.95	19.23	32.45	11.36	44.44	22.11
GPT-4o (Achiam et al., 2024) +SoV (Ours)	ViT	**51.27**	**31.93**	**49.12**	**22.65**	**49.12**	**20.46**	**50.10**	**24.20**
GPT-4V (Achiam et al., 2023) +SoV (Ours)	ViT	**60.91**	**41.96**	**53.21**	**22.82**	**50.00**	**18.97**	**55.33**	**28.69**

4 Experiments

4.1 Models and Settings

We do not need to train any models for our method. We evaluate the model's performance in a zero-shot manner using VLLMs. We include both commercial models such as GPT-4V-turbo (Achiam et al., 2023)[1] and GPT-4o-2024-05-13 (Achiam et al., 2024) as well as open-sourced models including MiniGPT-4-Vicuna (Zhu et al., 2023)[2], LLaVA-1.5-7B (Liu et al., 2023)[3], Video-LLaVA-7B (Zhang et al., 2023a)[4].

[1]https://chatgpt.com/
[2]https://github.com/Vision-CAIR/MiniGPT-4
[3]https://github.com/haotian-liu/LLaVA
[4]https://github.com/PKU-YuanGroup/Video-LLaVA

4.2 Dataset details

We collect original images from ABC News website[5]. Following the collection, we undertake meticulous preprocessing, initially removing any identical and blurry images through deduplication. To minimize human effort and cost in data annotation, we employ DeepFace (Serengil and Özpınar, 2024) for emotion annotation. Subsequently, two human annotators revise and refine the image labels. Finally, to finalize the labels, we involved a third annotator who has a professional background in psychology to verify correctness of facial expressions with their domain knowledge. This procedure guarantees the quality of the annotated data used to construct benchmarks. Table 4 in appendix A.1 presents the dataset details used for testing a

[5]https://www.abc.net.au/news/

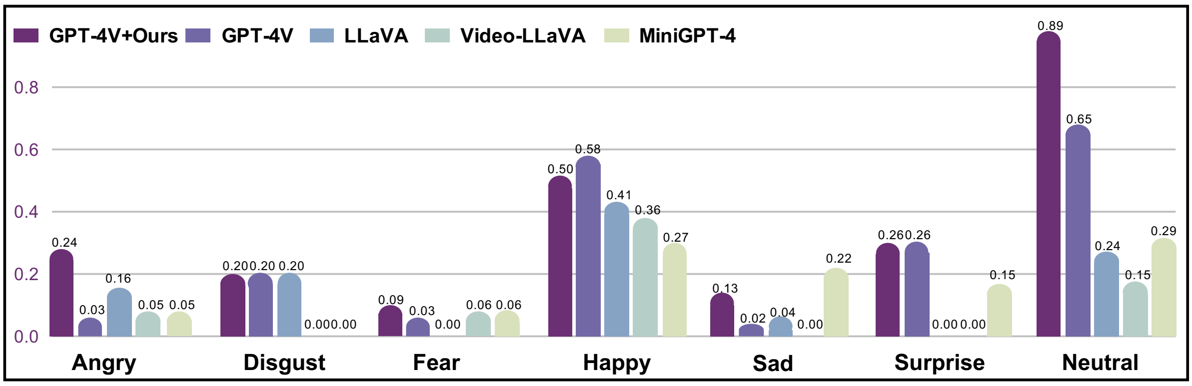


Figure 6: The bar chart shows the performance of various VLLMs in recognizing different emotions from images. The models compared include GPT-4V+Ours, GPT-4V (Achiam et al., 2023), LLaVA (Liu et al., 2023), Video-LLaVA (Zhang et al., 2023a), and MiniGPT-4 (Zhu et al., 2023). These results are distributed across seven emotions.

model on the task of zero-shot emotion recognition, structured across three different levels of difficulty: Easy, Medium, and Hard. This structured dataset aids in understanding the robustness and adaptability of the model in varying conditions of visual complexity.

4.3 Quantitative Results

Table 1 provides a detailed analysis of different zero-shot emotion recognition methods. MiniGPT-4 (Zhu et al., 2023) exhibits low performance, with accuracy ranging from 15.78% to 30.45% and Recall@1 from 12.85% to 16.17%. LLaVA (Liu et al., 2023) and Video-LLaVA (Zhang et al., 2023a) perform better in simpler categories but struggle significantly in the Hard category, where accuracy plummets to 3.50% and Recall@1 to 1.58%. In contrast, GPT-4V (Achiam et al., 2023) demonstrates robust performance across all levels, markedly improved by the SoV prompts. Specifically, GPT-4V+SoV achieves an impressive 60.91% accuracy and 41.96% Recall@1 in the Easy category, maintaining 50.00% accuracy and 18.97% Recall@1 even in the Hard category. These results underline SoV's effectiveness in boosting the model's ability to accurately interpret emotions across different complexities.

In Fig. 6, the chart highlights the varied efficacy of different VLLMs in emotion recognition tasks. GPT-4V+Ours consistently outperforms other models across nearly all emotions, particularly excelling in neutral and angry emotions. This highlights the specialized capabilities of GPT-4V+Ours in capturing more nuanced and varied emotional states. Meanwhile, other models show selective strengths and general weaknesses, particularly in recognizing negative emotions like fear and disgust.

4.4 Visual Prompting

Table. 2 presents a comparative analysis of state-of-the-art methods for zero-shot emotion recognition across datasets categorized by varying levels of difficulty: Easy, Medium, and Hard. It details performance metrics such as Accuracy and Recall, comparing the effectiveness of different visual prompting strategies utilized by GPT-4V. The baseline method (GPT-4V), which uses plain text prompts, demonstrates moderate effectiveness, with an overall accuracy of 44.44% and a Recall of 22.11%. Methods such as ReCLIP, RedCircle, and SoV employ more complex combinations of visual prompts. SoV (Ours), incorporating Numbers, Boxes, and Facial Landmarks, achieves the highest overall accuracy and recall scores of 55.33% and 28.69%, respectively. This suggests that the integration of multiple visual cues, particularly those that enhance the recognition of facial features, significantly improves performance across all difficulty levels, especially in more challenging datasets.

Fig. 7 visually represents how each visual prompting approach modifies the image to focus on emotion-relevant features. ReCLIP and RedCircle blur out non-facial areas and highlight faces with rectangles and circles, respectively. SoV applies a combination of visual prompts to emphasize facial areas while maintaining background context, which is critical for emotion recognition.

Fig. 8 reveals a comparative analysis of the performance across different emotional categories using three different visual prompts for GPT-4V: SoV (Ours), RedCircle, and ReCLIP. In this assessment,

Table 2: **Comparison of SOTA methods for zero-shot emotion recognition across datasets with varying levels of difficulty—Easy, Medium, and Hard.** The types of visual prompts used by previous approaches are: **P**: Crop, **B**: Box, **R**: Blur Reverse, **C**: Circle, **N**: Number, **F**: Facial Landmarks.

SOTA methods	Visual Prompt	Easy		Medium		Hard		Total	
		Acc (%)	R@1	Acc (%)	R@1	Acc (%)	R@1	Acc (%)	R@1
Baseline (Achiam et al., 2023)	Plain Text	48.85	27.94	47.95	19.23	32.45	11.36	44.44	22.11
ReCLIP (Subramanian et al., 2022)	P | B | R	**54.02**	**31.47**	46.19	16.98	42.10	14.32	48.14	22.98
RedCircle (Shtedritski et al., 2023)	P | C | R	51.72	29.55	**48.53**	**23.19**	**45.61**	**15.89**	**49.01**	**23.89**
SoV (Ours)	N | B | F	**60.91**	**41.96**	**53.21**	**22.82**	**50.00**	**18.97**	**55.33**	**28.69**

Figure 7: Visualization of the SOTA visual prompting approaches such as ReCLIP (Subramanian et al., 2022), RedCircle (Shtedritski et al., 2023) and our SoV prompts. ReCLIP and RedCircle blur out non-facial areas and highlight faces with rectangles and circles, respectively. SoV(Ours) applies a combination of visual prompts to emphasize facial areas while maintaining background context.

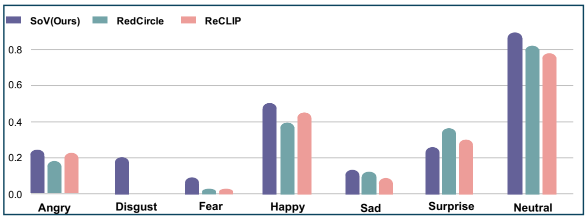


Figure 8: The bar chart illustrates the performance of SoV(Ours), RedCircle and ReCLIP in emotion recognition across seven different emotional categories.

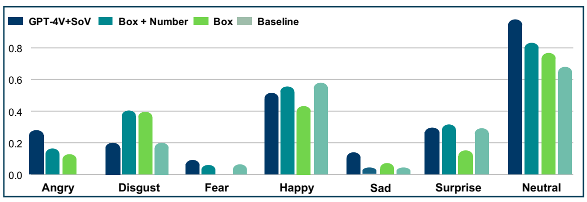


Figure 9: The bar chart displayed in the image illustrates the performance of different vision prompts—GPT-4V+SoV, Box + Number, Box, Baseline in emotion recognition across seven different emotional categories.

SoV outperforms the other methods in most emotional categories. Notably, SoV scores exceptionally well for 'Happy' and 'Neutral' emotions, suggesting a robust capability in recognizing these emotions accurately. RedCircle and ReCLIP show similar performance for 'Fear' and 'Sad', but lag significantly behind SoV in 'Happy' and 'Neutral'. For 'Angry', SoV and ReCLIP perform almost equally well, both surpassing RedCircle. This chart underscores SoV's strengths in nuanced emotion recognition, particularly in distinguishing positive and neutral states effectively.

4.5 Ablation Study

We investigate the effects of different vision prompts on GPT-4V. The data from Table 3 demonstrates how various vision prompts affect an emotion recognition system's performance across different difficulty levels. Baseline prompt shows moderate effectiveness with an accuracy of 48.85% in Easy, 47.95% in Medium, but drops to 32.45% in Hard scenarios. The overall accuracy is 44.44%, with Recall@1 similarly distributed, suggesting it performs consistently across different complexities but struggles with harder categories. Introduction of bounding boxes slightly reduces accuracy in simpler categories but improves Recall@1, suggesting enhanced precision in pinpointing relevant emotions. Notably, combining box and number prompts markedly boosts performance across all categories, particularly in challenging environments where accuracy rises significantly from 32.45% to 42.10% and Recall@1 from 11.36% to 15.57%. The SoV prompt outperforms all other methods, achieving peak accuracies of 60.91% in Easy and 50.00% in Hard, along with the highest

Table 3: **Ablation study for vision prompts on GPT-4V. Baseline**: represents the model's performance without any additional prompts. **Box**: indicates a visual prompt that uses bounding boxes. **Box+Number**: adding numerical identifiers to the bounding boxes. **SoV**: adding facial landmarks to each face with additional numerical identifiers to the bounding boxes.

Vision Prompt	Easy		Medium		Hard		Total	
	Acc (%)	R@1	Acc (%)	R@1	Acc (%)	R@1	Acc (%)	R@1
Baseline (Achiam et al., 2023)	48.85	27.94	47.95	19.23	32.45	11.36	44.44	22.11
Box	47.12	29.47	45.61	17.73	39.47	12.46	44.66	23.52
Box+Number	**58.04**	**41.10**	**51.46**	**22.12**	**42.10**	**15.57**	**51.63**	**28.24**
SoV	**60.91**	**41.96**	**53.21**	**22.82**	**50.00**	**18.97**	**55.33**	**28.69**

Figure 10: The impacts of segmentation masks for emotion recognition. **Top:** SoV provides a clearer view for emotion recognition. **Bottom:** the segmentation masks obscure parts of their faces, making it more challenging to accurately discern these emotions, especially for Person 2. In addition, the added segmentation masks also result in a lack of precise context.

Recall@1 figures, confirming its effectiveness in accurately interpreting emotions even in the most difficult settings.

Fig. 9 showcases an ablation study that compares the effectiveness of different vision prompts on GPT-4V. The GPT-4V+SoV configuration consistently outperforms the other methods in nearly all emotional categories, particularly excelling in 'Neutral' and 'Angry'. While the Box + Number prompts demonstrates moderate success in 'Happy' and 'Surprise', it still falls short compared to SoV prompts in other emotion categories. This result highlights that after adding extra facial landmarks, the VLLMs can capture more nuanced and varied emotional states.

4.6 Qualitative Observations

The image in Fig. 10 shows three people at a picnic setting, each displaying happy and relaxed expression, which suggests they are enjoying the outing. However, due to the segmentation masks applied, the masks obscure parts of their faces, making it more challenging to accurately discern these emotions, especially for person 2. In addition, the added segmentation masks also result in a lack of precise context.

5 Limitations

We suggest the implementation of Set-of-Vision prompting to bridge visual and textual prompts. However, a challenge arises as it is difficult to precisely describe visual prompts, such as the shape, location, or color of a bounding box and facial landmarks, in language. This issue might require encoding the visual prompts and fine-tuning the entire model for better accuracy. Moreover, this method is computationally intensive, potentially limiting its scalability and practicality in real-time applications, especially when handling large datasets or streaming videos such as tracking one person's emotion in different video frames.

6 Conclusion

In conclusion, our Set-of-Vision prompting (SoV) approach significantly advances the field of emotion recognition within VLLMs by addressing critical challenges in spatial localization and global context preservation. By leveraging spatial information such as bounding boxes and facial landmarks, SoV enhances zero-shot emotion recognition accuracy, ensuring precise face count and emotion categorization. Our face overlap handling algorithm and combined text-vision prompting strategy further refine the recognition process, highlighting the efficacy of integrating visual prompts in VLLMs for more accurate and detailed emotion analysis. This approach not only preserves the enriched image context but also offers a solution for detailed and nuanced emotion recognition, underscoring its potential impact on various applications within computer vision and natural language processing.

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Openai. *ChatGPT-4o https://www.openai.com/chatgpt*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sidong Feng and Chunyang Chen. 2024. Prompting is all you need: Automated android bug replay with large language models. In *Proceedings of the IEEE/ACM International Conference on Software Engineering*, pages 1–13.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Sefik Serengil and Alper Özpınar. 2024. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi*, 17(2):95–107.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11987–11997.

Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Transactions on Visualization and Computer Graphics*.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. 2022. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5198–5215.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2024. Vllms provide better context for emotion understanding through common sense reasoning. *arXiv preprint arXiv:2404.07078*.

Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. 2023a. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023b. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2024. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Sitao Zhang, Yimu Pan, and James Z Wang. 2023b. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18993–19004.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Proceedings of the Conference on Neural Information Processing Systems*, 36.

A Appendix

A.1 Dataset details

In the dataset Table 4, the "easy" dataset includes pictures with three or fewer faces. The "medium" dataset includes pictures with 3 to 7 faces. The "hard" dataset includes pictures with more than 7 faces. Easy dataset contains 76 images with a total of 174 faces. Medium dataset consists of 34 images featuring 171 faces. Hard dataset is the smallest set, comprising 9 images but still containing a significant number of faces (114). The table categorizes the datasets based on the complexity and density of faces in the images, which likely affects the challenge level for the model's emotion recognition capabilities. The usage of SoV prompts across all categories suggests a consistent testing approach, aiming to evaluate how well the model can interpret and predict emotions without prior specific training on these images (zero-shot learning). The metrics, Accuracy and Recall, are chosen to assess the model's precision in correctly identifying emotions and its ability to retrieve relevant instances across the datasets, respectively.

Table 4: The table presents the dataset details used for testing a model on the task of zero-shot emotion recognition, structured across three different levels of difficulty: Easy, Medium, and Hard.

Dataset	#Images	#Faces	Prompts	Metrics
Easy	76	174	SoV	Accuracy & Recall
Medium	34	171	SoV	Accuracy & Recall
Hard	9	114	SoV	Accuracy & Recall
Total	119	459	SoV	Accuracy & Recall

A.2 Scene based emotion recognition

The image in Fig. 11 captures a moment from a sports event, specifically a match between teams from Australia and the Philippines, with Australia leading based on the scoreboard ("PHI 0 - 1 AUS"). The scene includes various spectators and players, each showing distinct emotions which can aid in understanding the context and overall sentiment related to the ongoing game. Including scene context can significantly enhance the performance of language and image models (LLMs) in recognizing emotions. By understanding not just the facial expressions but also the situational context (such as the score in a sports game), models can make more accurate inferences about the probable emotions being displayed. In addition, environmental cues like scoreboards, team colors, and body language provide additional data points that help in accurately deducing the emotional state of individuals in group settings. Our SoV approach can bridge the gap between purely facial expression-based recognition and a more situation-aware understanding, leading to more nuanced and accurate emotion recognition capabilities in VLLMs.

In Fig. 12, the SoV method, by focusing on specific individuals within the image and retaining the clarity of the background context, offers a comprehensive approach to emotion recognition. This allows for a detailed analysis of not just the visible facial expressions, but also the situational context and interactions among individuals, providing a more nuanced understanding of the emotions being conveyed. For example, in the case of Person 2, despite a partially visible face, the SoV approach leverages additional contextual cues to deduce the emotion more accurately. In contrast, the RedCircle approach primarily relies on clear facial expressions by blurring background details, potentially overlooking subtler emotional cues embedded in the environment and body language.

A.3 Chain of thought for emotion recognition

The "Chain of Thought" approach, as illustrated in Fig. 13, offers advantages by breaking down the thought process into sequential, logical steps. This method enhances accuracy in interpreting emotions by considering various contextual clues such as activities, and body language, which are then systematically analyzed. For example, the analysis first identifies the activity (Halloween trick-or-treating) and setting, which sets the emotional backdrop. Subsequently, it classifies each individual based on visible attributes and costumes, leading to more nuanced emotion recognition. This step-by-step reasoning mirrors human cognitive processes, allowing for more refined and contextually appropriate interpretations of emotions, such as distinguishing neutral expressions from smiles, even in a complex social setting like Halloween, where expressions might otherwise be ambiguous.

Scene based Emotion Recognition

User: Is the team with the red cloth girl winning or losing? What should I say to her? What's emotion for the girl label "1" and '2' and '3', and '4' ?

GPT-4V + SoV:

Answer:

The scoreboard in the image shows **"PHI 0 - 1 AUS,"** indicating that the team associated with **Australia** (presumably the team in light blue) is **winning**, and the team associated with the **Philippines** (possibly the team the girl in red is supporting, if she is a fan or player of this team) is **losing**.

If you wish to speak to the girl in red and she is supporting or part of the Philippine team, here are a few supportive things you could say:

1. Encouraging her spirit: "The game's not over yet! There's still time for a comeback."

2. Showing empathy and support: "It's a tough game, but you all are showing great spirit! Keep cheering!"

3. Acknowledging the effort: "It's been a challenging match, but I admire the team's resilience. How are you holding up?"

These comments are supportive and considerate, acknowledging the situation while remaining **positive and encouraging**.

Based on their expressions in the image:

1. The girl **labeled "1"** seems focused and somewhat intense, likely reflecting her **determination** in the game given that her team is **currently losing**.

2. The girl **labeled "2"** appears **joyful and is smiling**, which suggests she is **happy**, likely because her team is in the **lead**.

3. The person **labeled "3"** is **also smiling** and appears to be **enjoying the moment**, which is consistent with a positive emotion like **happiness or contentment**.

4. The person **labelled "4"** (in the inset) has a neutral expression, which might suggest he **is calm or composed**, possibly **observing the game with interest** or concentration.

Figure 11: Scene based Emotion Recognition

Compare RedCircle and SoV prompts

RedCircle | **SoV**

User: What's emotion for these person?

GPT-4V + RedCircle :

Answer:

In the image you've shared, there are several people showing different emotions:

1.The man in the blue shirt is smiling widely, indicating happiness or joy.

2.The girl to his left also seems happy, with a broad smile.

3.The woman in the foreground on the right is smiling as well, suggesting she is also feeling happy.

4.The boy in the foreground on the left appears to be laughing or speaking enthusiastically, which generally reflects excitement or happiness.

Overall, the group appears to be enjoying a pleasant and happy moment together.

GPT-4V + SoV:

Answer:

In this image set within a home dining scene:

1.Person 1 (woman in purple square): She shows a joyful and welcoming expression, likely enjoying the conversation and the meal.

2.Person 2 (boy in red square): He looks attentive and possibly listening or thinking deeply about the conversation. His expression suggests engagement rather than emotional display.

3.Person 3 (man in green square): He is smiling as he eats, reflecting happiness and satisfaction, perhaps with the meal and the company.

4.Person 4 (girl in blue square): She also shows a happy demeanor, smiling broadly, indicating she is having a good time.

Overall, the scene conveys a warm and cheerful family mealtime, where everyone seems to be enjoying both the food and each other's company.

Figure 12: Compare RedCircle and SoV prompts for emotion recognition.

Chain of thought for emotion recognition

User:
Step 1: What are they planning to do?
Step 2: How many people are there in the image?
Step 3: Can you give me the related labels for these person?
Step 4: What are their emotions?

GPT-4V + SoV:

Question 1:

The group in the image seems to be participating in a Halloween trick-or-treating event. They are dressed in costumes and carrying buckets for collecting candy, which is a traditional activity where participants visit houses to ask for treats. The setting, with residential cars and houses in the background, further supports that they are likely walking through a neighborhood to gather sweets from different homes.

Question 2:

There are four people in the image.

Question 3:

In the image:

1. Person 1 (on the far right) - labeled as "1", wearing a black shirt with white skeleton bones printed on it and carrying an orange pumpkin bucket.
2. Person 2 (second from the right) - labeled as "2", dressed in casual clothes, holding documents and an orange bucket.
3. Person 3 (second from the left) - labeled as "3", dressed in a black tutu and carrying an orange bucket.
4. Person 4 (on the far left) - labeled as "4", also dressed in a black tutu, but with a blue bucket instead.

Each label corresponds to the numbered box drawn over their faces in the image.

Question 4:

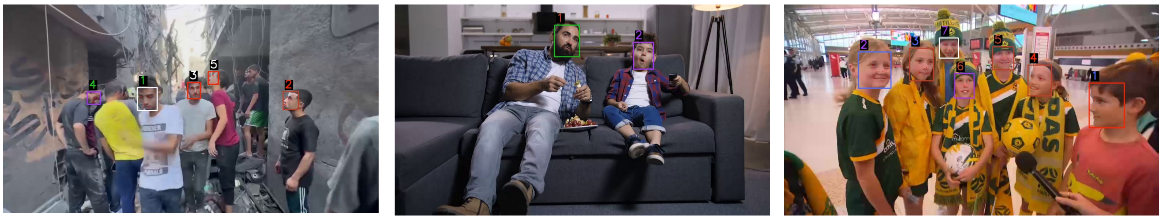
From the visible expressions in the image:

1. Person 1 (on the far right): Appears to have a neutral expression.
2. Person 2 (second from the right): Their face is not visible in the image.
3. Person 3 (second from the left): Seems to be smiling, indicating happiness or cheerfulness.
4. Person 4 (on the far left): Appears to have a neutral expression.

These are the inferred emotions based on the limited facial expressions visible in the image.

Figure 13: Chain of thought for emotion recognition

Detect negative emotions for group of people



User:
My friend Sally mentioned seeing a very sad news video that deeply affected her. Which of the images could she have seen?

GPT-4V + SoV:
If Sally mentioned seeing a very sad news video that deeply affected her, it's likely she referred to the image depicting people in a scene of destruction or after a disaster. This would be the first image of the three presented in your query, where individuals are seen amidst rubble and damage. Such scenes typically involve serious issues like natural disasters, conflicts, or other tragic events that can have a profound emotional impact on viewers. This type of content aligns with a description of a "very sad news video."
The other two images show more casual or celebratory scenarios that are less likely to be described as deeply affecting in a sad context.

User:
What are the face emotion of the people in the first image?

GPT-4V + SoV:
In the image where individuals are amidst rubble, the emotions of the people can be described as follows:

1. **Person 1**: Appears to be in distress or discomfort, potentially indicative of shock or sadness.
2. **Person 2**: Looks serious and concerned, likely reflecting the gravity of the situation.
3. **Person 3**: His expression is hard to discern fully due to the angle, but he seems to be focused and possibly worried.
4. **Person 4**: Shows a pensive or somber expression, which might suggest contemplation or distress.
5. **Person 5**: Seems to have a neutral or tired expression, which might reflect fatigue or resignation in the face of the situation.

Overall, the expressions suggest a mixture of concern, sadness, and seriousness, aligning with a response to a serious or tragic event.

Figure 14: Detect negative emotions for different people

Grounding ability

User: If you are a reporter and want to ask a question about the game, which labeled person is the best person for you to talk to?

GPT-4V + SoV:

Answer:

In the image you provided from a sporting event, the person labeled "2" would be the most appropriate individual for a reporter to ask questions about the game. This person is likely a coach or player involved in the game, given that he is in uniform and holding a child who is also in a similar uniform. This suggests he is directly related to the team and would have insights about the game's specifics.

User: what is the emotion of the person labeled "2"

GPT-4V + SoV:

Answer:

The person labeled "2" in the image appears to be smiling and engaging positively with the interviewer, which suggests that he is feeling happy or content during the interview. His overall demeanor and facial expression convey a sense of enjoyment or satisfaction, likely related to the context of the sports event.

Figure 15: Grounding ability