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†]

Australian National University[1] Quriosity Pty Ltd[2] Webumate Pty Ltd[3] Curtin University[4] Yale University[5]

Introduction

- The tasks of emotion recognition requires the decoding of emotions from nuanced indicators like facial expressions, body language, and contextual details.
- Previous methods overlook the spatial relationships between different people and facial features within a single face.
- The relationships between the eyes, mouth, and nose features can be highlighted by facial landmarks in SoV prompts to guide VLLMs.
- Recent approaches approaches focus on local objects and ignore spatial context information.

Objectives

- We Introduce 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 the need for cropping faces, thus maintaining the holistic view of the image.
- The proposed face overlap handling algorithm effectively addresses conflicts arising from overlapping face detections, especially in images with dense face clusters.
- Our results show that incorporating spatial visual prompts (SoVs) into VLLMs can enhance their performance in recognizing emotions.

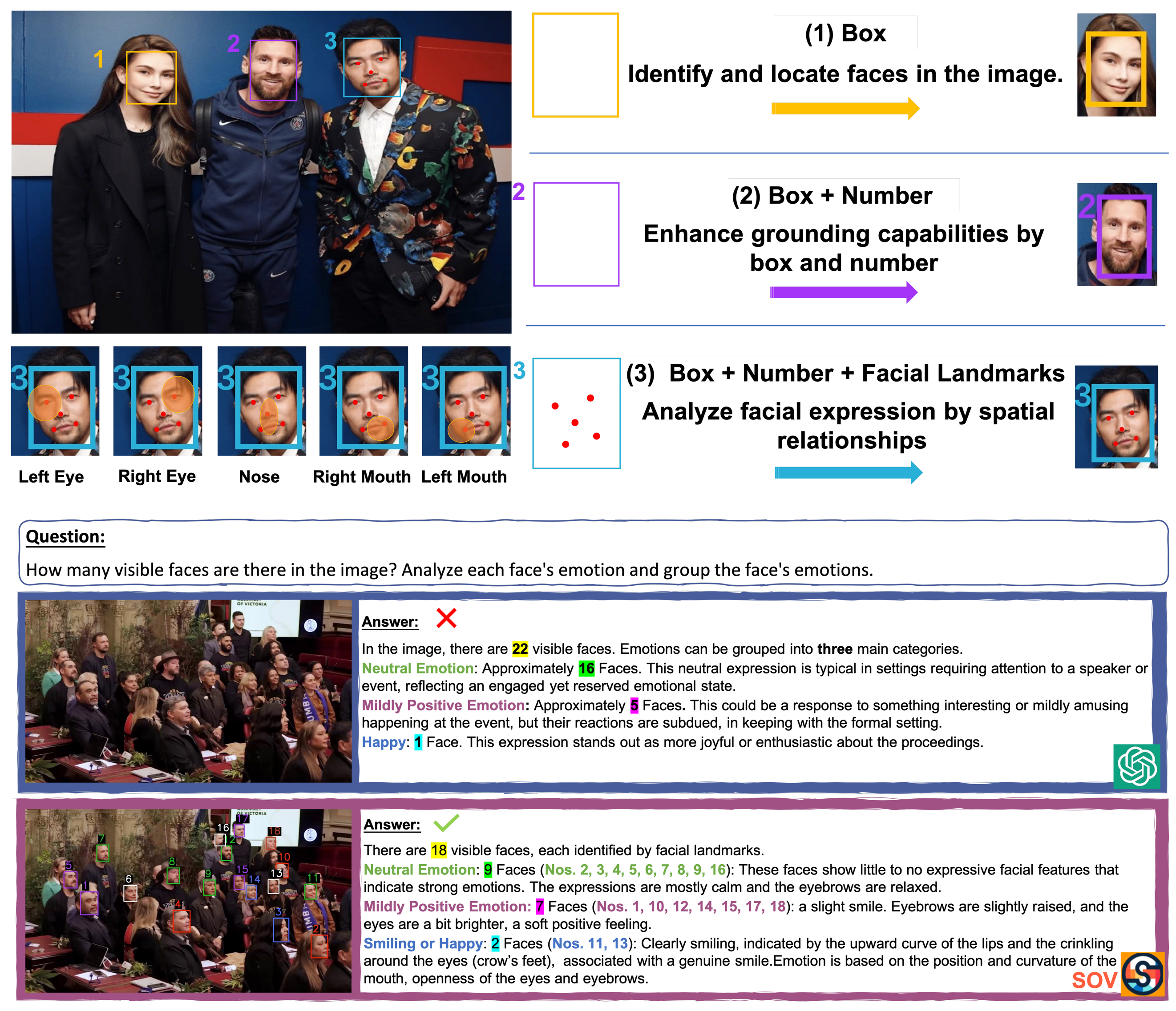

Figure 5: We use two types of prompt methods. **Left**: plain text prompts, which can be used for group emotion recognition. **Right**: combined text-vision prompts, which can be used for analyzing specific individuals' emotions. These prompts can be used to evaluate emotional interpretation in relation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

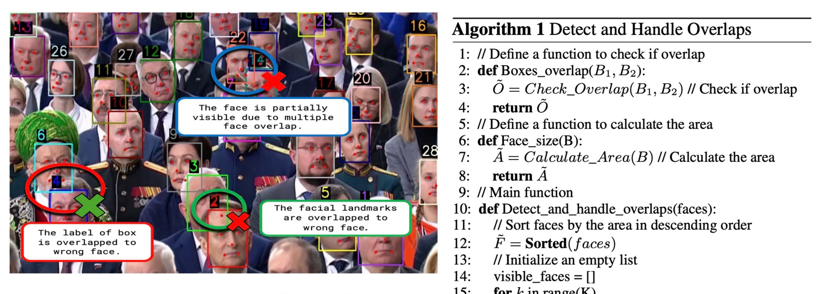

Figure 4: **Face detection inevitably introduces some overlaps or conflicts that confuse VLLMs.** Analyzing the impact of face overlaps, occlusions, landmark misalignment, and bounding box conflicts for emotion recognition.

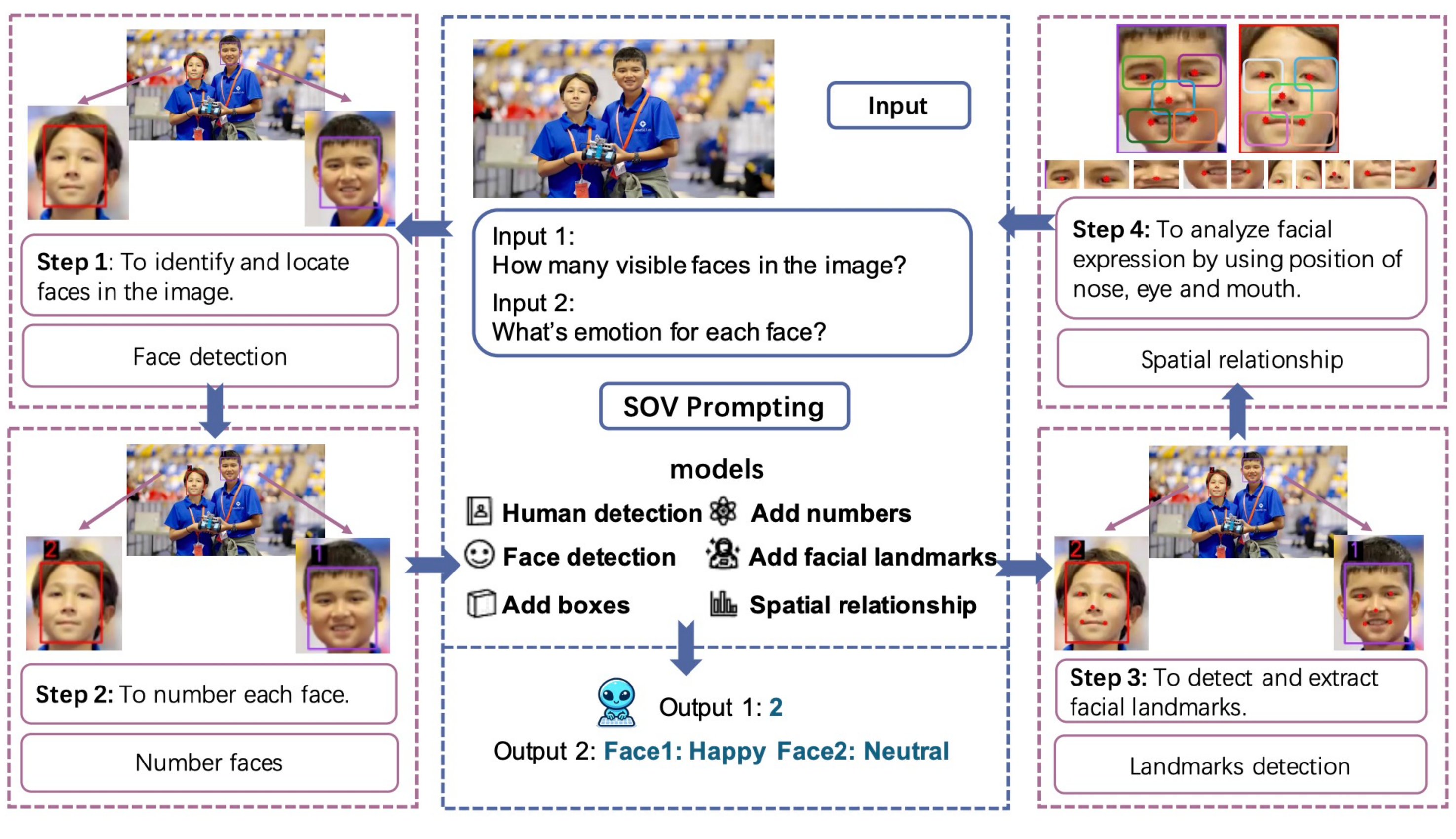

Figure 3: **Workflow diagram for enhanced face recognition and emotion analysis using the Set-of-Vision (SoV)** prompting approach: a multi-step process involving face detection, face numbering, landmark extraction, and spatial relationship analysis for emotion classification. Each detected face is analyzed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

Methodology

Result

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

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).

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 I B I R	54.02	31.47	46.19	16.98	42.10	14.32	48.14	22.98
RedCircle (Shtedritski et al., 2023)	P I C I R	51.72	29.55	48.53	23.19	45.61	15.89	49.01	23.89
SoV (Ours)	N I B I F	60.91	41.96	53.21	22.82	50.00	18.97	55.33	28.69

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.

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

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.

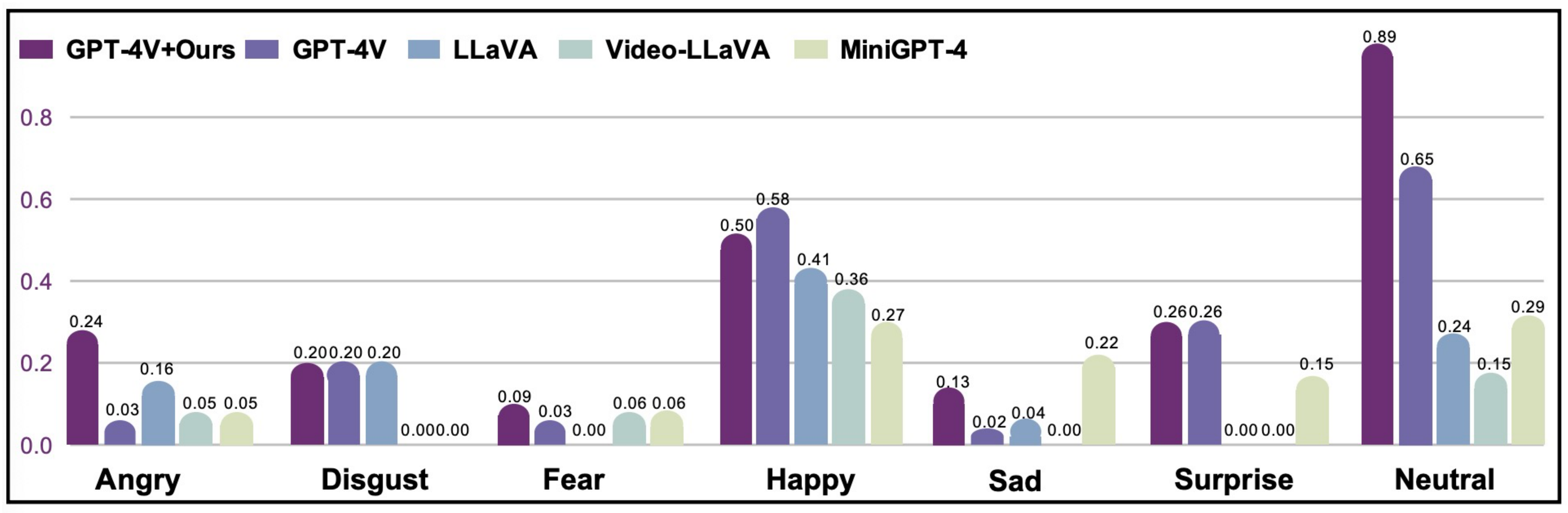

Figure 6: The bar chart shows the performance of various VLLMs in recognizing different emotions from images. The models compared include GPT-4V+Ours, GPT-4V (Achiam et al., 2023), LLaVA (Liu et al., 2023), Video-LLaVA (Zhang et al., 2023a), and MiniGPT-4 (Zhu et al., 2023). These results are distributed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

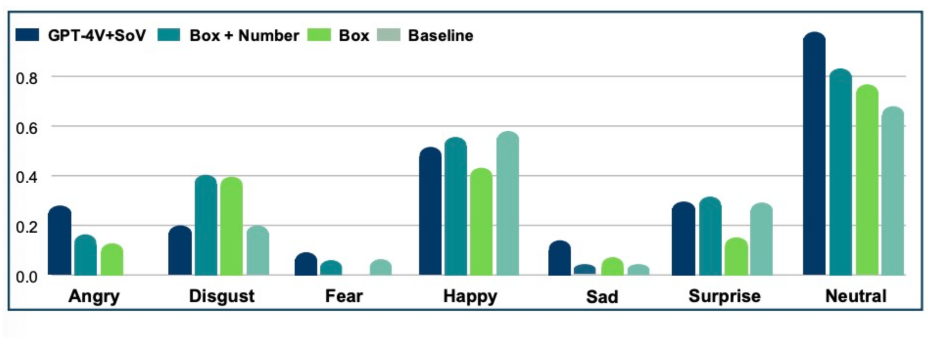

Figure 9: The bar chart displayed in the image illustrates the performance of different vision prompts—GPT-4V+SoV, Box + Number, Box, Baseline in emotion recognition across seven different emotional categories.

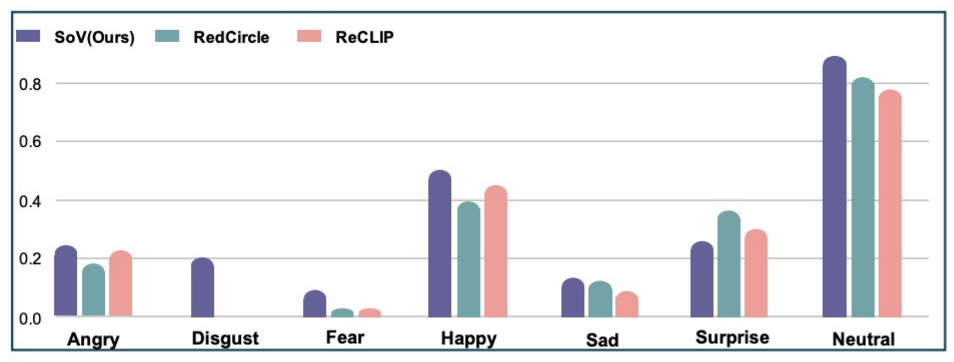

Figure 8: The bar chart illustrates the performance of SoV(Ours), RedCircle and ReCLIP in emotion recognition across seven different emotional categories.

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.

Conclusion

- ✓ Face overlap handling algorithm and combined text-vision prompting strategy further refine the recognition process.
- ✓ This approach not only preserves the enriched image context but also offers a solution for detailed and nuanced emotion recognition.
- ✓ Set-of-Vision prompting (SoV) approach significantly advances the field of emotion recognition within VLLMs.
- ✓ SoV enhances zero-shot emotion recognition accuracy, ensuring precise face count and emotion categorization.