

1. Case Study Fot HECO

Figure 1 2 3 4 displays the CoT results on HECO. We compared the content generated by VILA in zero-shot, 2-shot, and 4-shot settings. We found that a single label is often insufficient to express a person’s emotions. For example, the person in the red box in Figure 4 is correctly predicted as ”Sadness” in both the 2-shot and 4-shot settings, but the zero-shot setting’s prediction of ”Disgusted” also reasonably fits the task. Of course, the model also makes completely incorrect inferences, such as Figure 1, in the 4-shot setting, the person is predicted with an entirely opposite emotion. While the description might seem logical when reading the text alone, examining the image reveals that the model’s description of the image is correct, but its understanding is flawed. This indicates that the model’s capability to comprehend images needs further improvement.

2. Limitation

Although our research demonstrates that LVLMs perform excellently and have great potential for further development in the CAER task, they still have the following limitations:

- Unlike LLMs, there are not many LVLMs with strong ICL capabilities. This is mainly because most LVLMs have not been trained with multi-image data, making it difficult for them to understand demonstrations with images during inference. Our work relies on the ICL capabilities of LVLMs, which limits our choice of models. In the future, we can try converting demonstration images into textual descriptions. This would not only help LVLMs better understand the content but also reduce the input sequence length, allowing LVLMs to process more demonstrations.
- The performance of LVLMs is highly dependent on the original parameter size and the scale of the training data. In this paper, we only explored the performance of models with fewer than 10 billion parameters. Verifying whether larger LVLMs would perform better on the CAER task requires additional computational resources.

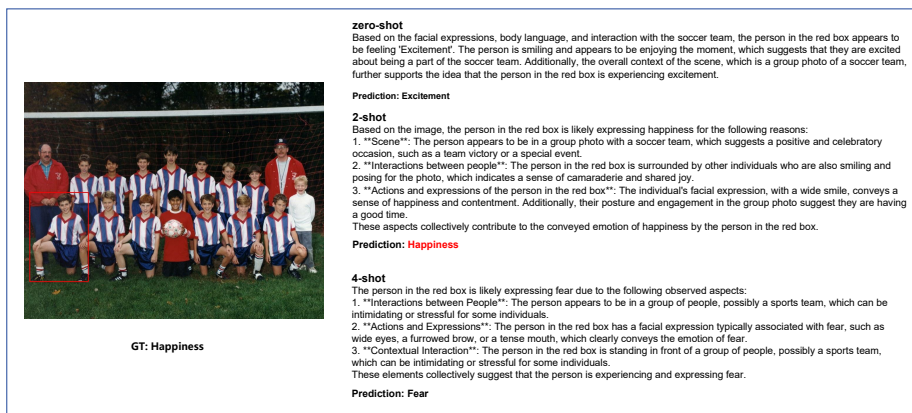


Figure 1: Examples of LVLm’s CoT results on HECO dataset.

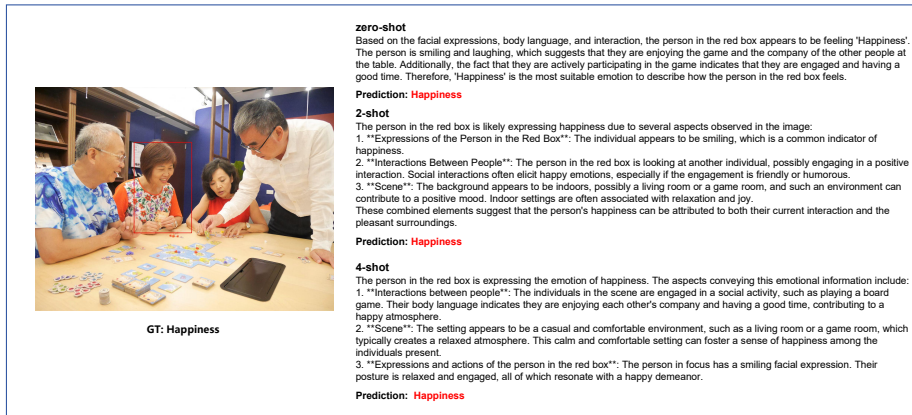


Figure 2: Examples of LVLML's CoT results on HECO dataset.

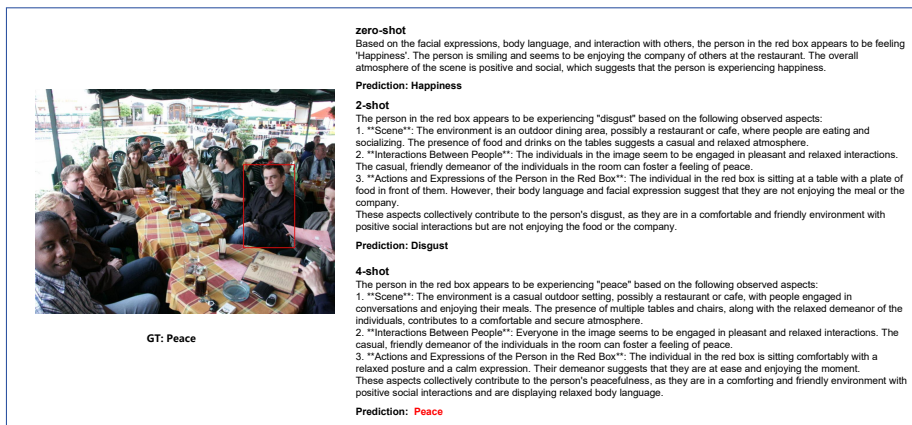


Figure 3: Examples of LVLML's CoT results on HECO dataset.

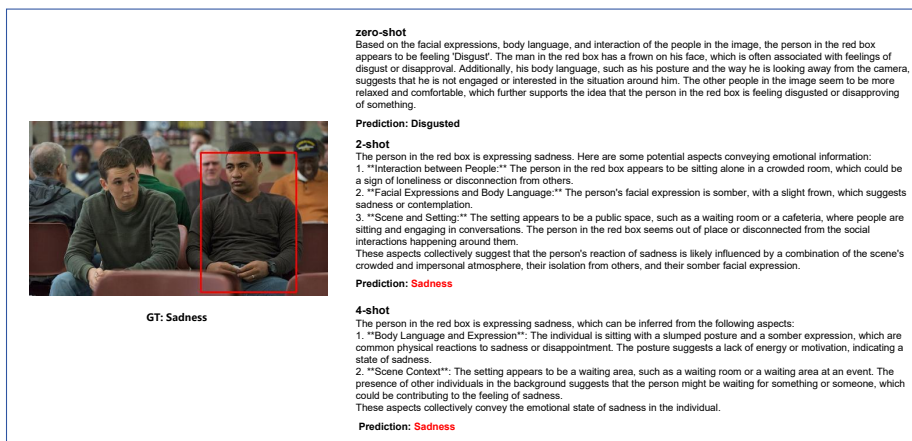


Figure 4: Examples of LVLML's CoT results on HECO dataset.