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.

GT: Happiness

zero-shot
Based on the facial expressions, body language, and interaction with the soccer team, the person in the red box appears to be feeling 'Excitement'. The person is smiling and appears to be enjoying the moment, which suggests that they are excited about being a part of the soccer team. Additionally, the overall context of the scene, which is a group photo of a soccer team, further supports the idea that the person in the red box is experiencing excitement.

Prediction: Excitement

2-shot
Based on the image, the person in the red box is likely expressing happiness for the following reasons:
1. **Scene**: The person appears to be in a group photo with a soccer team, which suggests a positive and celebratory occasion, such as a team victory or a special event.
2. **Interactions between people**: The person in the red box is surrounded by other individuals who are also smiling and posing for the photo, which indicates a sense of camaraderie and shared joy.
3. **Actions and expressions of the person in the red box**: The individual's facial expression, with a wide smile, conveys a sense of happiness and contentment. Additionally, their posture and engagement in the group photo suggest they are having a good time.
These aspects collectively contribute to the conveyed emotion of happiness by the person in the red box.

Prediction: Happiness

4-shot
The person in the red box is likely expressing fear due to the following observed aspects:
1. **Interactions between People**: The person appears to be in a group of people, possibly a sports team, which can be intimidating or stressful for some individuals.
2. **Actions and Expressions**: The person in the red box has a facial expression typically associated with fear, such as wide eyes, a furrowed brow, or a tense mouth, which clearly conveys the emotion of fear.
3. **Contextual Interaction**: The person in the red box is standing in front of a group of people, possibly a sports team, which can be intimidating or stressful for some individuals.
These elements collectively suggest that the person is experiencing and expressing fear.

Prediction: Fear

Figure 1: Examples of LVLM's CoT results on HECO dataset.

GT: Happiness

zero-shot
Based on the facial expressions, body language, and interaction, the person in the red box appears to be feeling 'Happiness'. The person is smiling and laughing, which suggests that they are enjoying the game and the company of the other people at the table. Additionally, the fact that they are actively participating in the game indicates that they are engaged and having a good time. Therefore, 'Happiness' is the most suitable emotion to describe how the person in the red box feels.

Prediction: Happiness

2-shot
The person in the red box is likely expressing happiness due to several aspects observed in the image:
1. **Expressions of the Person in the Red Box**: The individual appears to be smiling, which is a common indicator of happiness.
2. **Interactions Between People**: The person in the red box is looking at another individual, possibly engaging in a positive interaction. Social interactions often elicit happy emotions, especially if the engagement is friendly or humorous.
3. **Scene**: The background appears to be indoors, possibly a living room or a game room, and such an environment can contribute to a positive mood. Indoor settings are often associated with relaxation and joy.
These combined elements suggest that the person's happiness can be attributed to both their current interaction and the pleasant surroundings.

Prediction: Happiness

4-shot
The person in the red box is expressing the emotion of happiness. The aspects conveying this emotional information include:
1. **Interactions between people**: The individuals in the scene are engaged in a social activity, such as playing a board game. Their body language indicates they are enjoying each other's company and having a good time, contributing to a happy atmosphere.
2. **Scene**: The setting appears to be a casual and comfortable environment, such as a living room or a game room, which typically creates a relaxed atmosphere. This calm and comfortable setting can foster a sense of happiness among the individuals present.
3. **Expressions and actions of the person in the red box**: The person in focus has a smiling facial expression. Their posture is relaxed and engaged, all of which resonate with a happy demeanor.

Prediction: Happiness

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

GT: Peace

zero-shot
Based on the facial expressions, body language, and interaction with others, the person in the red box appears to be feeling 'Happiness'. The person is smiling and seems to be enjoying the company of others at the restaurant. The overall atmosphere of the scene is positive and social, which suggests that the person is experiencing happiness.

Prediction: Happiness

2-shot
The person in the red box appears to be experiencing "disgust" based on the following observed aspects:
1. **Scene**: The environment is an outdoor dining area, possibly a restaurant or cafe, where people are eating and socializing. The presence of food and drinks on the tables suggests a casual and relaxed atmosphere.
2. **Interactions Between People**: The individuals in the image seem to be engaged in pleasant and relaxed interactions. The casual, friendly demeanor of the individuals in the room can foster a feeling of peace.
3. **Actions and Expressions of the Person in the Red Box**: The individual in the red box is sitting at a table with a plate of food in front of them. However, their body language and facial expression suggest that they are not enjoying the meal or the company.
These aspects collectively contribute to the person's disgust, as they are in a comfortable and friendly environment with positive social interactions but are not enjoying the food or the company.

Prediction: Disgust

4-shot
The person in the red box appears to be experiencing "peace" based on the following observed aspects:
1. **Scene**: The environment is a casual outdoor setting, possibly a restaurant or cafe, with people engaged in conversations and enjoying their meals. The presence of multiple tables and chairs, along with the relaxed demeanor of the individuals, contributes to a comfortable and secure atmosphere.
2. **Interactions Between People**: Everyone in the image seems to be engaged in pleasant and relaxed interactions. The casual, friendly demeanor of the individuals in the room can foster a feeling of peace.
3. **Actions and Expressions of the Person in the Red Box**: The individual in the red box is sitting comfortably with a relaxed posture and a calm expression. Their demeanor suggests that they are at ease and enjoying the moment.
These aspects collectively contribute to the person's peacefulness, as they are in a comforting and friendly environment with positive social interactions and are displaying relaxed body language.

Prediction: Peace

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

GT: Sadness

zero-shot
Based on the facial expressions, body language, and interaction of the people in the image, the person in the red box appears to be feeling 'Disgust'. The man in the red box has a frown on his face, which is often associated with feelings of disgust or disapproval. Additionally, his body language, such as his posture and the way he is looking away from the camera, suggests that he is not engaged or interested in the situation around him. The other people in the image seem to be more relaxed and comfortable, which further supports the idea that the person in the red box is feeling disgusted or disapproving of something.

Prediction: Disgusted

2-shot
The person in the red box is expressing sadness. Here are some potential aspects conveying emotional information:
1. **Interaction between People:** The person in the red box appears to be sitting alone in a crowded room, which could be a sign of loneliness or disconnection from others.
2. **Facial Expressions and Body Language:** The person's facial expression is somber, with a slight frown, which suggests sadness or contemplation.
3. **Scene and Setting:** The setting appears to be a public space, such as a waiting room or a cafeteria, where people are sitting and engaging in conversations. The person in the red box seems out of place or disconnected from the social interactions happening around them.
These aspects collectively suggest that the person's reaction of sadness is likely influenced by a combination of the scene's crowded and impersonal atmosphere, their isolation from others, and their somber facial expression.

Prediction: Sadness

4-shot
The person in the red box is expressing sadness, which can be inferred from the following aspects:
1. **Body Language and Expression**: The individual is sitting with a slumped posture and a somber expression, which are common physical reactions to sadness or disappointment. The posture suggests a lack of energy or motivation, indicating a state of sadness.
2. **Scene Context**: The setting appears to be a waiting area, such as a waiting room or a waiting area at an event. The presence of other individuals in the background suggests that the person might be waiting for something or someone, which could be contributing to the feeling of sadness.
These aspects collectively convey the emotional state of sadness in the individual.

Prediction: Sadness

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