Can Multimodal LLMs Learn to Better Recognize Emotions?

Jared Guerrero Moreno, Leticia Pinto-Alva (PhD Mentor), Prof. Jesse Thomason (PI)

Computer Science Department, Viterbi School of Engineering, University of Southern California

Introduction

In emergencies such as disasters, it is hard to find survivors and identify if they are in distress. Autonomous search robots can help, but they might have trouble identifying if a person is in distress if no external injuries are observable. In this case, a Machine Learning (ML) model capable of recognizing emotions would aid in determining if a person is in distress, or in good condition. Accordingly, our objective is to make a ML model which can recognize emotions better than current models. To make our model, we will use LLAVA [1], a Multimodal LLM (Large Language Model), as the base. Then, we will fine-tune LLaVA on the EMOTIC dataset [2], which has images of people in everyday context with annotations on their emotions, to make our model with better emotion recognition.

Data

The EMOTIC dataset has 18K images where each person in the images is annotated from a list of 26 emotions as seen in **Fig. 1**. Additionally, this list of emotions is also accompanied by Valence, Arousal, and Dominance (**VAD**) scores. These scores will help the model learn to recognize emotions better since each score represents different aspects of a person's emotional state.

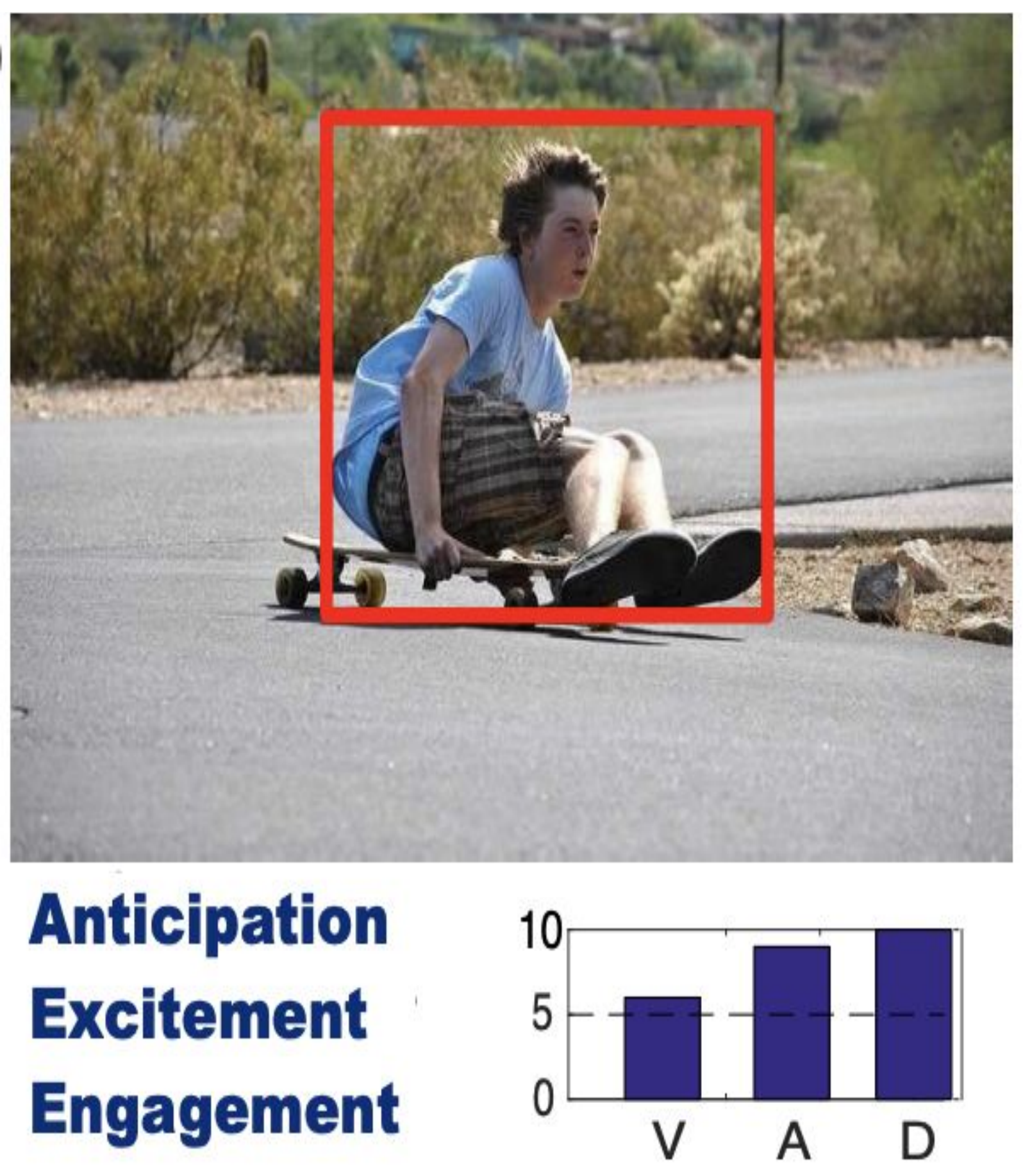
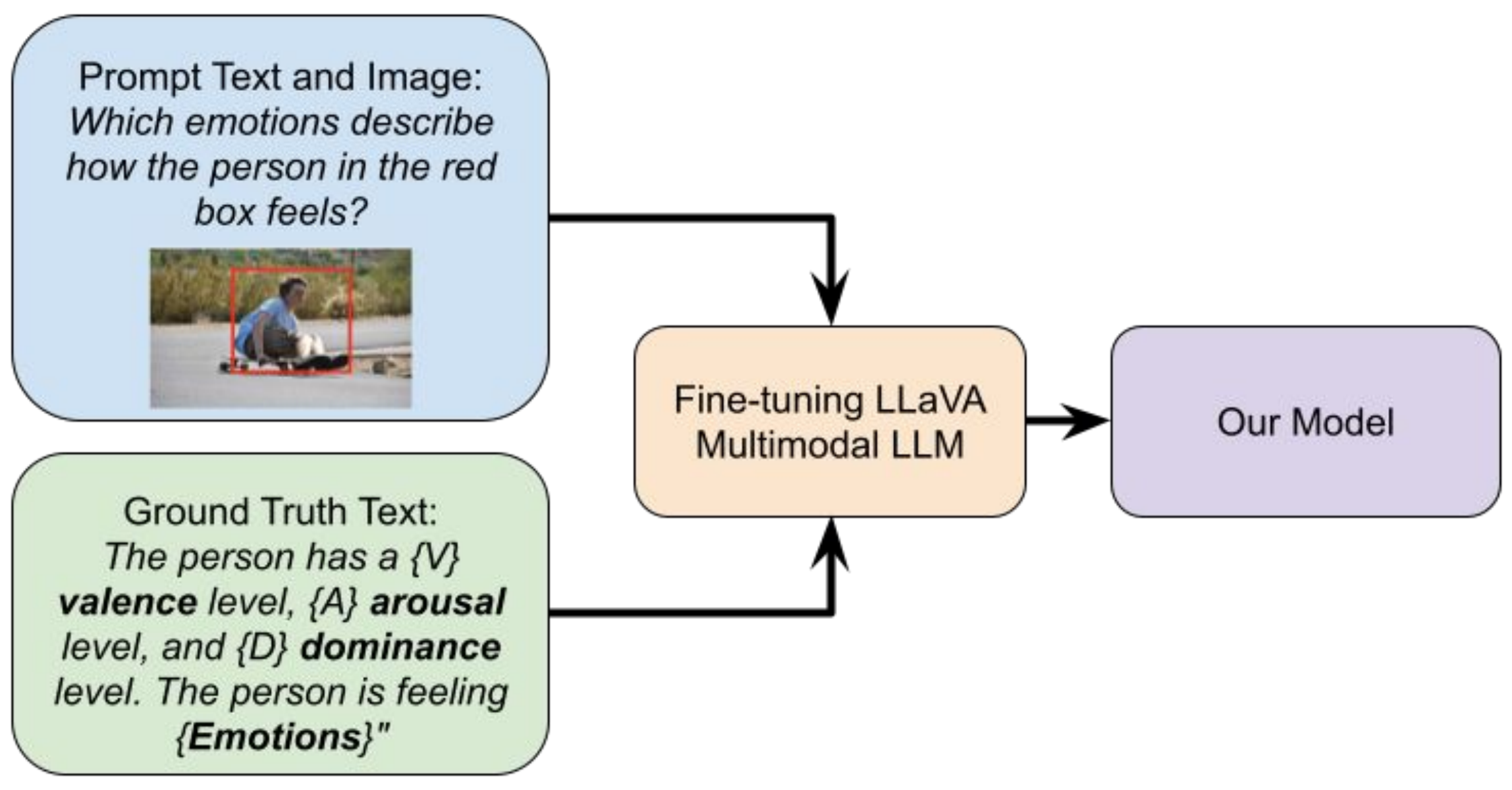


Figure 1: Image from EMOTIC dataset with annotation

More specifically**,** **Valence** measures how positive an emotion the person is feeling, while **Arousal** measures the person's agitation level, and **Dominance** measures the control level of the situation by the person.

Methods



Prompt Text and Image: *Which emotions describe how the person in the red box feels?*

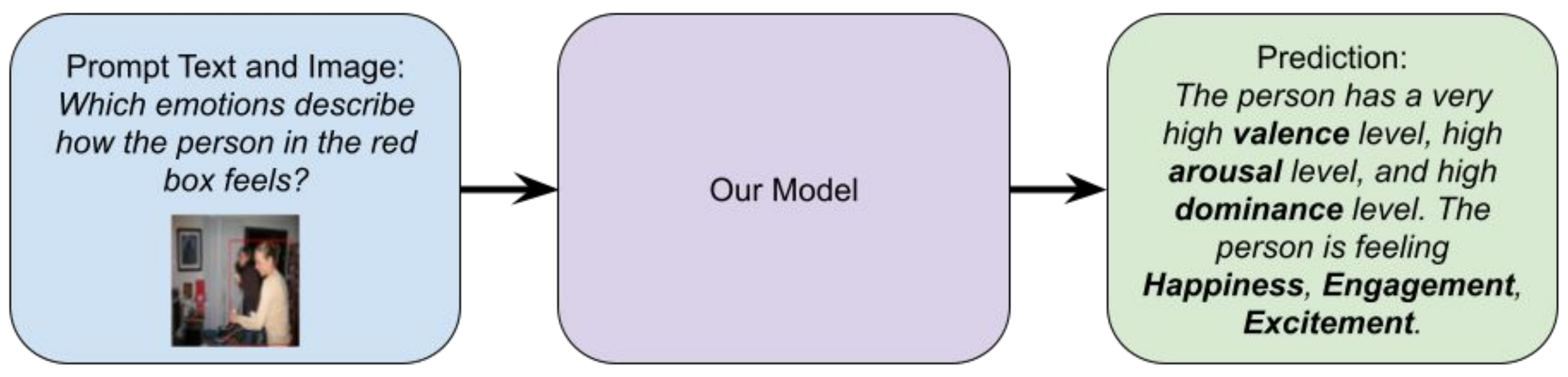
Ground Truth Text: *The person has a {V}* ***valence*** *level, {A}* ***arousal*** *level, and {D}* ***dominance*** *level. The person is feeling {****Emotions****}"*

Fine-tuning LLaVA Multimodal LLM

Our Model

Figure 2: Workflow for fine-tuning LLaVA to make our model

We fine-tuned LLaVA with images from the EMOTIC dataset as shown in **Fig. 2** to make our model. The ground truth consists of the VAD scores converted to text ranging from *very low* (1) to *very high* (10), as well as all the annotated emotions within the Emotions field. Our model was then tested as seen in **Fig. 3**. Since we used both emotions and VAD scores in the ground truth, we performed an ablation study where we repeated these steps with the exception of ground truth consisting of only the Emotions field, without any VAD scores.



Prompt Text and Image: *Which emotions describe how the person in the red box feels?*

Our Model

Prediction: *The person has a very high* ***valence*** *level, high* ***arousal*** *level, and high* ***dominance*** *level. The person is feeling* ***Happiness*, *Engagement*, *Excitement***.

Figure 3: Workflow for testing our model

Results

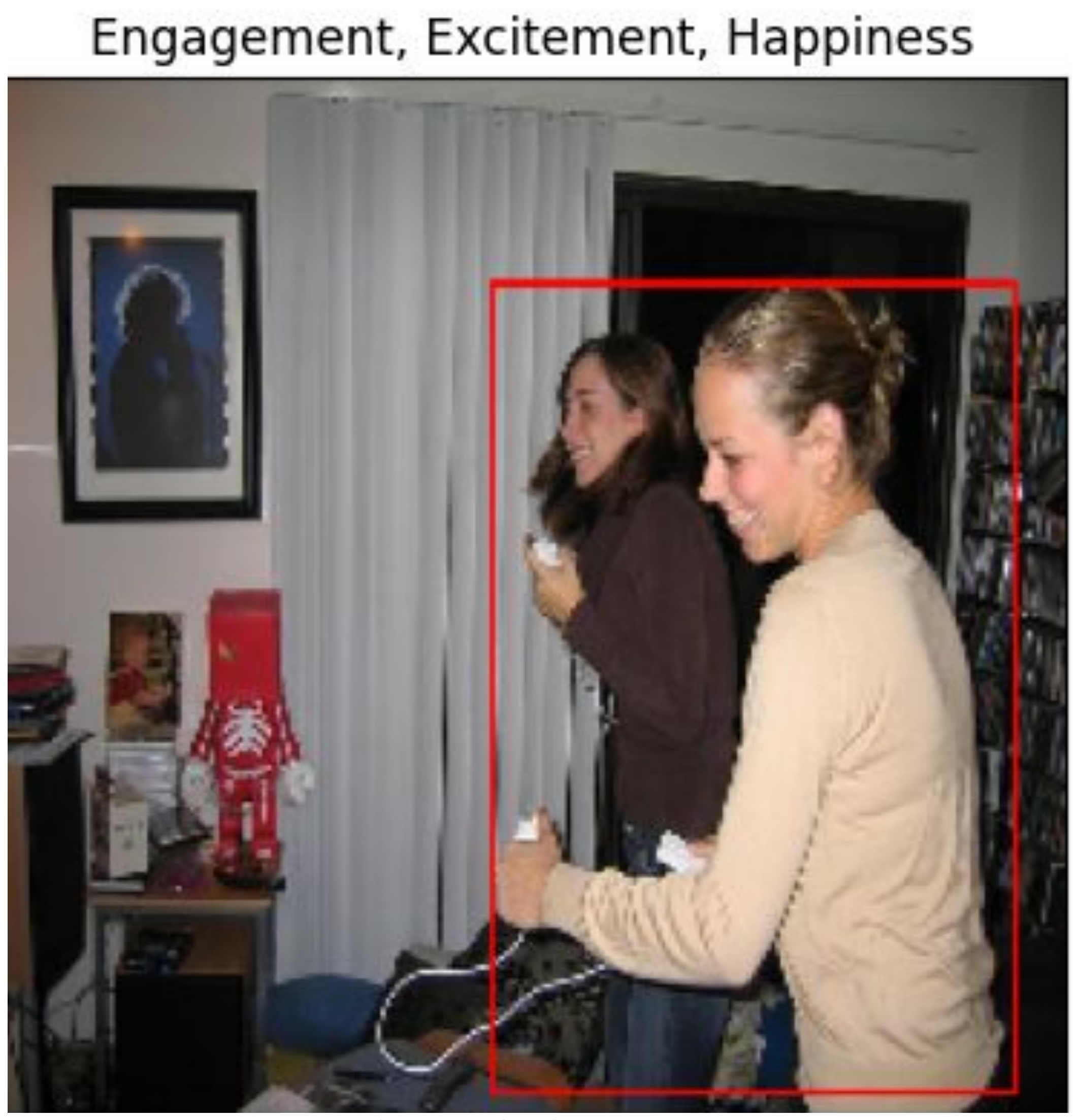


Figure 4: Image from EMOTIC dataset with our predictions

Model	mAP↑
LLaVA Fine-tuned with Emotion List & VAD Scores	**40.60**
VLLMs Provide Better Context for Emotion Understanding Through Common Sense Reasoning* [3]	38.52
LLaVA Fine-tuned with Emotion List Only	32.71
LLaVA (zero-shot)*	16.98

Table 1:Comparison to state-of-the-art on EMOTIC
* These metrics were provided by [3]

Table 1 shows our model outperforms the current SOTA for Emotion Recognition by 2% on mAP (mean Average Precision), which measures the performance of a model for tasks such as object detection tasks and information retrieval.

The ground truth emotions in **Fig. 4** are ***Engagement***, ***Excitement***, and ***Happiness***. As seen in **Fig. 3**, our model correctly predicts ***Happiness***, ***Engagement***, and ***Excitement***. For comparison, LLaVA's zero-shot output is: "*The person in the red box is likely feeling* ***happy*** *as they are smiling while playing the video game.*"

In contrast to those models, our model fine-tuned solely on emotions predicts: ***Happiness, Excitement.*** Our ablation study results seen in **Figs. 4** and **5** show that our model achieves a higher mAP and accuracy when fine-tuned on both emotions and VAD scores rather than just emotions. Furthermore, our results show that our model reached its highest mAP and accuracy on the sixth epoch of the fine-tuning process.

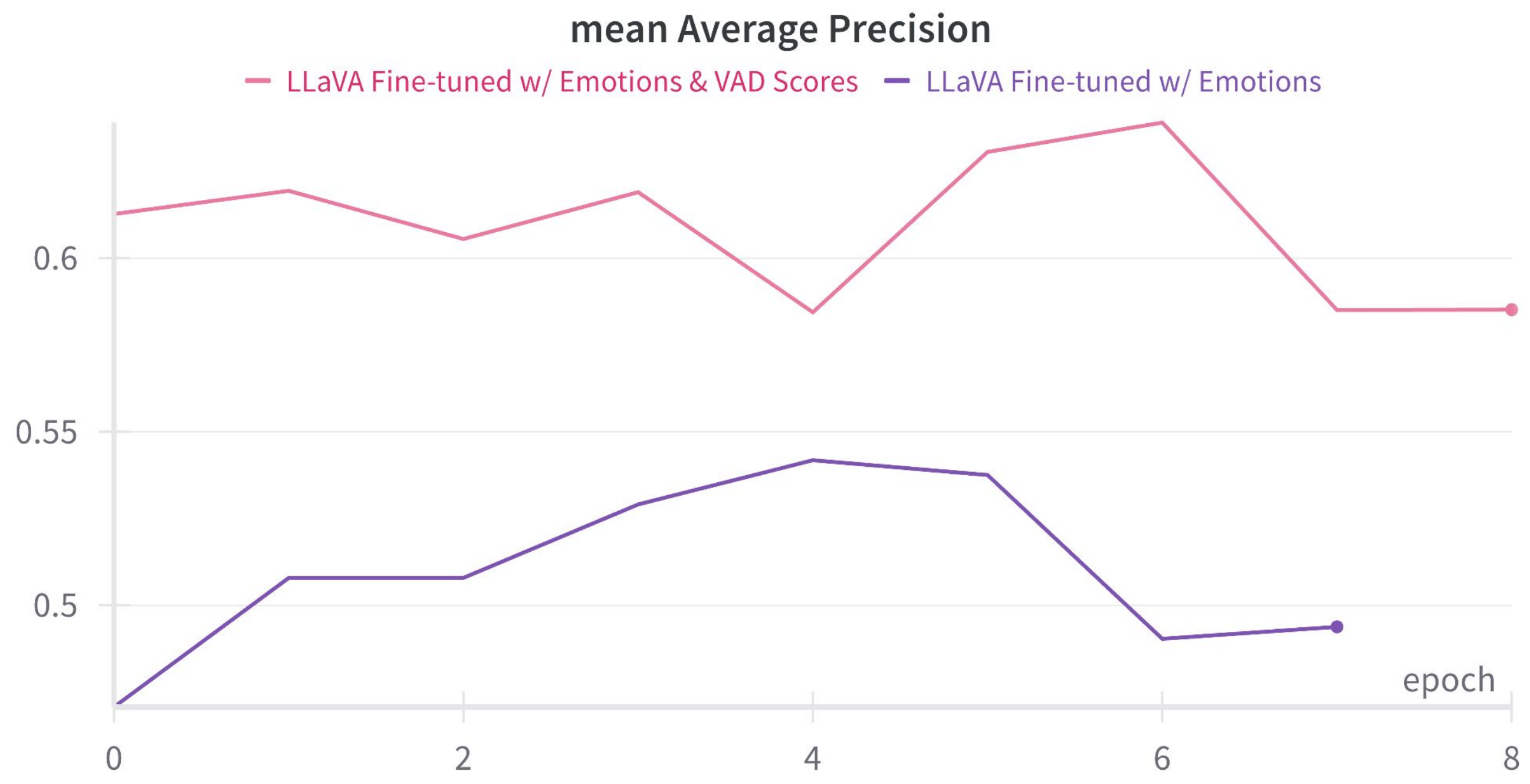


Figure 4: mAP comparison after fine-tuning

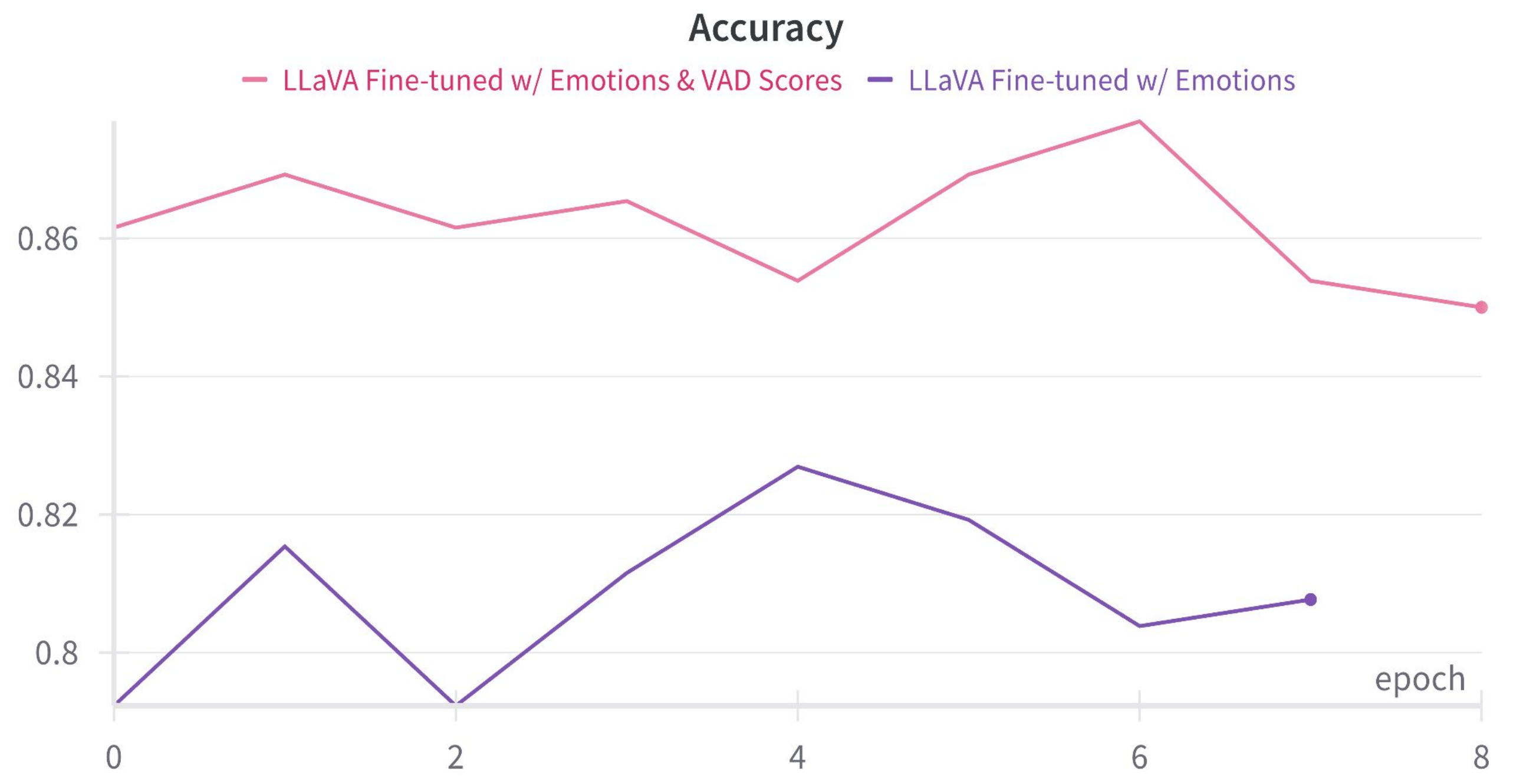


Figure 5: Accuracy comparison after fine-tuning

Conclusion

Our research suggests that fine-tuning a Multimodal LLM on the EMOTIC dataset with images and corresponding emotions as ground truth improves emotion recognition results. Finally, providing VAD scores as ground truth in addition to each image's corresponding emotions during fine-tuning further improves the model's ability to recognize emotions.

Next Steps

Our next steps include investigating if providing LLaVA with context as shown in [3] combined with our methods would result in better emotion recognition. Since LLaVA's VQA capabilities are desired to describe the environment a person is in, research into mitigating catastrophic forgetting during fine-tuning is necessary. In line with our motivation of aiding people in emergencies, a curated image dataset of people in distress with annotations similar to those found in the EMOTIC dataset would be useful to make.

Acknowledgements

This research project was made possible thanks to the University of Southern California's (USC) GLAMOR Lab, Prof. Jesse Thomason, and Leticia Pinto Alva, whose mentorship and guidance were particularly helpful. This research project was funded in part by Amazon through USC's Viterbi SURE (Summer Undergraduate Research Experience) program.

References

1) Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. (2023)
2) Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In CVPR, July 2017.
3) Xenos, Alexandros, et al. "VLLMs Provide Better Context for Emotion Understanding Through Common Sense Reasoning." arXiv preprint arXiv:2404.07078 (2024).