# Exploring Emotion Recognition from Images through Vision-Language Models: A Case Study on LLaVA

CRISTINA-MARIA TOADER, University of Twente, The Netherlands

With the rapid advancement of artificial intelligence and the increasing frequency of human-AI interaction, the ability of AI systems to recognize human emotions has become an important area of research. Despite extensive research conducted on the topic, achieving a human-level understanding of the emotion present in an image is still an open problem, particularly in real-world scenarios involving complex social and contextual cues. Recent developments in vision-language models (VLMs) have introduced new opportunities by combining visual perception with linguistic reasoning, mimicking the way humans perceive and interpret emotions across diverse situations. These models have demonstrated promising results, in some cases outperforming traditional emotion recognition approaches.

## 1 INTRODUCTION

### 1.1 Background and Context

For many years, emotion recognition from images using computer vision techniques focused primarily on facial analysis [1, 15], leveraging detailed facial features, and employing Convolutional Neural Networks (CNNs) to classify emotions. While these approaches can achieve high accuracy in controlled settings, their performance often deteriorates in real-world environments. Psychological studies have demonstrated that, in addition to facial expressions, the context and surroundings play a crucial role in accurately interpreting emotions [9]. As a result, several datasets have been developed to represent people in diverse, real-world situations and to facilitate research that incorporates contextual information. One of the first and most representative of these is the EMOTIC dataset [9]. The EMOTIC baseline model introduced a dual-stream architecture, separately processing the features of the target person and the surrounding scene using two CNNs, and then fusing these features to predict a set of discrete emotion categories and continuous affective dimensions. This approach outperformed models that relied solely on person-centered features, highlighting the importance of contextual information for accurate emotion recognition [9].

Following the success of the EMOTIC approach, further research has explored increasingly sophisticated architectures for incorporating contextual information into emotion recognition [5, 14, 15]. These efforts include the use of advanced CNN and DCNN (Deep Convolutional Neural Networks) variants [4], graph-based reasoning models [5], and, more recently, vision-language models (VLMs) [7], which are pretrained on image-text pairs to capture rich semantic and contextual cues.

### 1.2 Problem Statement

Despite extensive research in the field of emotion recognition from images, it is not yet clear which type of model or architecture is optimal for this task, as no existing approach consistently achieves high performance across diverse, real-world scenarios [7, 9]. Traditional models that rely solely on facial expressions struggle to capture the complex interplay between individuals and their surroundings and, while graph-based models have improved contextual reasoning, they are still limited by the completeness and accuracy of detected visual elements and are vulnerable to computer vision failures like occlusion, poor lighting, or pose ambiguity [10, 16, 17]. Furthermore, graph-based models that use image captions for emotion recognition are limited by the quality of these captions: if a caption lacks sufficient descriptive detail or omits key contextual or emotional information, the resulting emotion predictions are likely to be less accurate [5]. Moreover, such models are constrained by the limited vocabulary present in the training captions, which can lead to poor predictions when encountering words or expressions not seen during training. In this context, vision-language models (VLMs) appear as a possible solution for these problems, as they combine images and text, allowing them to better understand the meaning of a scene, compared to models that use only one type of input [2]. For this reason, the modern approach that uses VLMs has shown promise in the task of emotion recognition, challenging the previous models in terms of performance [7]. The full potential of VLMs is, however, not yet fully understood and continues to be an active area of research.

### 1.3 State-of-the-art in Emotion Recognition

For a long period of time, models achieved comparable performance for the task of emotion recognition, with an average mAP of 20–35% on the EMOTIC dataset (mean Average Precision: metric that summarizes both precision and recall across confidence thresholds) [11]. Recent advancements have significantly elevated the state-of-the-art; new combinations of deep convolutional neural networks (DCNNs), which integrate both person and scene context, have reached mAP scores of 79.60% and 78.39%, representing a substantial leap over earlier methods [11].

In parallel, the capabilities of vision-language models (VLMs) such as LLaVA [12], CLIP [18] and GPT-4 [6] are an active area of research in the task of emotion recognition, revealing that even zero-shot models, those not fine-tuned for this specific task, can achieve performance on par with traditional models from previous studies [7].

Alongside DCNNs and VLMs, graph-based models that use image captions for emotion prediction are also currently researched. One example of such a state-of-the-art model, also used in this study, is the Graph Isomorphism Network (GIN) developed by Costa et al. [5]. This model approaches emotion recognition by first generating

*TScIT 43, July 4, 2025, Enschede, The Netherlands*

captions for each image using ExpansionNet [8], a transformer-based image captioning model. These captions are then transformed into graph representations, which are used as input to the GIN to perform emotion classification. While this method achieved a respectable mAP of approximately 30% on the EMOTIC dataset, it has notable limitations. Firstly, because the original study uses ExpansionNet for image captioning, the generated captions primarily describe contextual elements of the scene and lack words related to emotions or facial expressions. This causes the model to miss important emotional cues. Secondly, the vocabulary of the model is limited to the words seen during training, making it less robust when encountering unfamiliar terms during testing.

In the present study, these limitations were addressed by replacing the ExpansionNet-generated captions with those produced by LLaVA, a state-of-the-art vision-language model [12]. LLaVA generates richer and more expressive captions that include emotion-related vocabulary and references to facial expressions.

Two state-of-the-art datasets in the field of emotion recognition, also used in this study, are EMOTIC [9] and FindingEmo [14].
The EMOTIC dataset is a large-scale collection of real-world images, where each person is annotated with a subset of 26 applicable discrete emotions and corresponding Valence-Arousal-Dominance (VAD) scores, enabling context-aware emotion recognition.
FindingEmo is a dataset of 25,869 real-world images, each annotated with a single emotion out of 24 possible labels, valence, and arousal score, capturing the overall sentiment of complex, multi-person scenes.

### 1.4 Aim of the Research

The goal of this research is to explore the capabilities of modern vision-language models, such as LLaVA [12], in the task of emotion recognition from images. More specifically, the desired objectives are the following:
**Goal 1:** To evaluate the accuracy of the state-of-the-art open-source vision-language model, LLaVA, in recognizing the emotions of an individual from images, and to compare it with that of a more traditional approach, namely the Graph Isomorphism Network (GIN) developed by Willams de Lima Costa [5]. This includes exploring how different image captioning strategies, used as input to the GIN, affect its performance.
**Goal 2**: To evaluate the accuracy of LLaVA in recognizing the general sentiment in images containing multiple individuals and complex, variable context.

### 1.5 The Research Questions

To accomplish the proposed goals, the research will be guided by the following overarching question, further divided into specific subquestions:
**Main Research Question:** What are the capabilities of the vision-language model LLaVA for the task of emotion recognition from images?
**RQ1:** How accurately can zero-shot LLaVA recognize all of the emotions felt by a person from images compared to the Graph Isomorphism Network (GIN) developed by Willams de Lima Costa, and how do different input captions affect the performance of the GIN?
**RQ2:** How accurately can LLaVA recognize the general sentiment of real-world images, with a variable number of subjects, as evaluated against the annotations in the FindingEmo dataset?

### 1.6 Contribution

This study evaluates the performance of the vision-language model LLaVA on two widely used emotion recognition datasets: EMOTIC and FindingEmo. For the FindingEmo dataset, LLaVA is also used to establish a zero-shot baseline in the Emo24 setting, where all 24 emotion labels are used for evaluation, as no prior benchmarks could be found for this full-label setup. In addition, the study compares the zero-shot performance of LLaVA to the Graph Isomorphism Network (GIN) introduced by Willams de Lima Costa [5]. Since in the original study the GIN was trained on captions generated by ExpansionNet, and in the present study, it is tested on LLaVA-generated captions, the study also provides insight into how testing the model on a different vocabulary than the one it was originally trained on can affect its performance.

## 2 GENERAL CONSIDERATIONS AND INITIAL EXPERIMENTS

### 2.1 General Considerations

The images shown in Figure 1, along with their corresponding labels, are chosen from the EMOTIC dataset and serve as a baseline reference for several of the examples discussed in the subsequent sections.

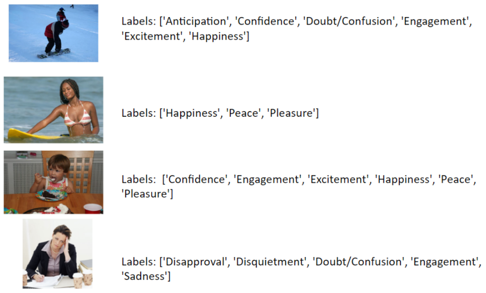

Labels: ['Anticipation', 'Confidence', 'Doubt/Confusion', 'Engagement', 'Excitement', 'Happiness']
Labels: ['Happiness', 'Peace', 'Pleasure']
Labels: ['Confidence', 'Engagement', 'Excitement', 'Happiness', 'Peace', 'Pleasure']
Labels: ['Disapproval', 'Disquietment', 'Doubt/Confusion', 'Engagement', 'Sadness']

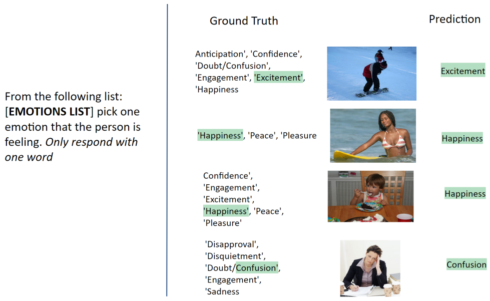
Fig. 1. Four images from the EMOTIC dataset their corresponding labels

Furthermore, the LLaVA model used for both image captioning and zero-shot emotion recognition evaluation was "llava-v1.5-7b".

For all the experiments that involved evaluating LLaVA predictions, both the labels and the outputs were first processed, prior to evaluation. This preprocessing involved converting all words to lowercase and applying stemming in order to ensure consistent matching with the ground truth labels. Stemming is the process of reducing words to their root form, allowing semantically similar variants such as "happy" and "happiness" to be treated as equivalent by reducing both to the common stem "happi".

### 2.2 Initial experiment

First, to obtain a better initial understanding of the capacities of LLaVA, the vision-language model was asked to identify the primary emotion experienced by the main person in each image, selecting a single emotion from a predefined list corresponding to the 26 categorical labels in the EMOTIC dataset. The obtained performance was then compared to the one of a random predictor. An example of the obtained predictions for the established set of images can be seen in Figure 2. This result was not used for comparison to the performance of the Graph Isomorphism Network, but rather to gain an initial insight into the capabilities of LLaVA.



From the following list: [**EMOTIONS LIST**] pick one emotion that the person is feeling. *Only respond with one word*

Ground Truth
Prediction
Anticipation', 'Confidence', 'Doubt/Confusion', 'Engagement', 'Excitement', 'Happiness
Excitement
'Happiness', 'Peace', 'Pleasure
Happiness
Confidence', 'Engagement', 'Excitement', 'Happiness', 'Peace', 'Pleasure'
Happiness
'Disapproval', 'Disquietment', 'Doubt/Confusion', 'Engagement', 'Sadness
Confusion

Fig. 2. The four example images, with their respective prediction and the prompt used for prediction generation

### *Implementation Details*

When evaluating single-emotion predictions against the multi-label EMOTIC annotations, accuracy was measured: a prediction was considered correct if the predicted emotion appeared in the list of annotated labels for that image.

To contextualize the performance of LLaVA, the accuracy of a random baseline was also included. The random generator was executed 200 times, and the highest accuracy obtained across these runs was reported in the results for comparison.

### *Results:*

Table 1 shows the accuracy of LLaVA in the task of single-emotion prediction on the EMOTIC dataset.

Table 1. Accuracy comparison of LLaVA and a random baseline for the task of single emotion prediction on the EMOTIC dataset

| **Model** | **EMOTIC dataset** |
|---|---|
| **Zero-shot LLaVA** | 60.88% |
| **Random Baseline***| 52.37% |

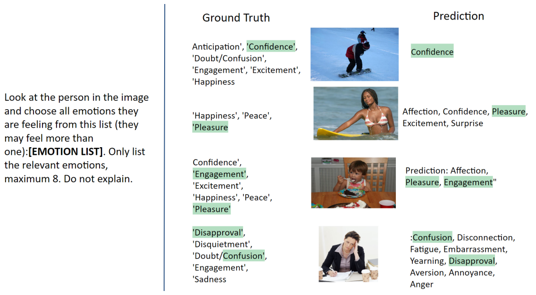
*highest obtained accuracy out of 200 trials

## 3 COMPARATIVE ANALYSIS OF LLAVA AND GIN ON CATEGORICAL EMOTION DETECTION

### 3.1 Methodology

The investigation used the EMOTIC dataset and comprised two main components: (a) evaluating the zero-shot emotion recognition performance of the vision-language model LLaVA, and (b) assessing the performance of the Graph Isomorphism Network (GIN) proposed by Willams de Lima Costa with different LLaVA-generated captions as input. The methodology also investigated the changes in performance of the GIN with different captions as input.

In order to assess component (a), LLaVA was prompted to provide a list of the emotions experienced by the target person in the image, selecting from the same predefined set of 26 EMOTIC emotion labels. An example of the multi-labels predictions can be seen in Figure 3. The output was evaluated using the micro F1 score, and was later used to compare the performance of the model against that of the Graph Isomorphism Network.



Look at the person in the image and choose all emotions they are feeling from this list (they may feel more than one):**[EMOTION LIST]**. Only list the relevant emotions, maximum 8. Do not explain.

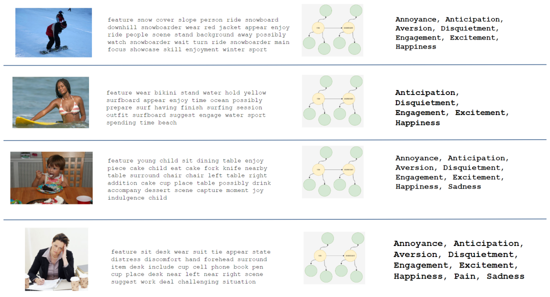
Ground Truth
Prediction
Anticipation', 'Confidence', 'Doubt/Confusion', 'Engagement', 'Excitement', 'Happiness
Confidence
'Happiness', 'Peace', 'Pleasure
Affection, Confidence, Pleasure, Excitement, Surprise
Confidence', 'Engagement', 'Excitement', 'Happiness', 'Peace', 'Pleasure'
Prediction: Affection, Pleasure, Engagement"
'Disapproval', 'Disquietment', 'Doubt/Confusion', 'Engagement', 'Sadness
:Confusion, Disconnection, Fatigue, Embarrassment, Yearning, Disapproval, Aversion, Annoyance, Anger

Fig. 3. The four example images, with their respective multi-label predictions and the prompt used for prediction generation

To understand component (b), a short overview of the method proposed by Willams de Lima Costa is provided further: The approach begins by generating a detailed caption for each provided image using an image captioning model. This caption undergoes several preprocessing steps: stop words and common nouns are removed, and the remaining words are lemmatized to produce a refined set of “valid words.” These words are then used to construct two co-occurrence matrices: one capturing how often each word appears with specific emotions, and another capturing how often words appear together. Semantic attributes for each word, such as mood tags, polarity, and related concepts, are retrieved using SenticNet [3]. This information is then used to build a graph where nodes represent words, emotions, and semantic descriptors, and edges reflect co-occurrence or semantic relationships. Finally, a Graph Isomorphism Network (GIN) is used to classify these graphs into one or more of the 26 EMOTIC emotion categories [5].

The same pipeline proposed by Willams de Lima Costa is employed in this research and is illustrated in Figure 4. The figure outlines the general process: input images are first processed into captions, which are then used to construct schematic graphs that serve as input to the GIN model, ultimately leading to the prediction of the corresponding emotions. The only difference between the original approach and the one employed in this research is that the image captions are generated using the vision-language model LLaVA, and the preprocessing steps, such as stemming, stopword removal, and filtering of common nouns, are performed manually instead of using the predefined methods from the original implementation. An example of the captions processing is shown in Figure 5. A comparison between the predictions made by the GIN and the annotations for the four example images is shown in Figure 6.

Furthermore, to assess how different input captions influence the predictions of the model, the performance of the GIN was also assessed with five different captions generated by LLaVA.



feature snow cover slope person ride snowboard downhill snowboarder wear red jacket appear enjoy ride people scene stand background away possibly watch snowboarder wait turn ride snowboarder main focus showcase skill enjoyment winter sport

Annoyance, Anticipation, Aversion, Disquietment, Engagement, Excitement, Happiness

feature wear bikini stand water hold yellow surfboard appear enjoy time ocean possibly prepare surf having finish surfing session outfit surfboard suggest engage water sport spending time beach

Anticipation, Disquietment, Engagement, Excitement, Happiness

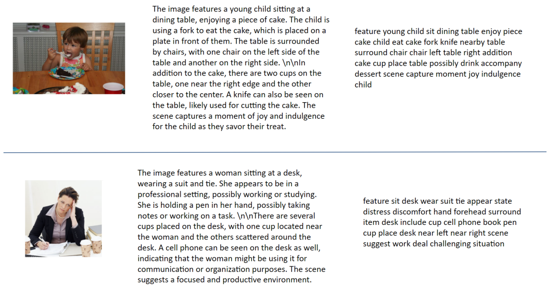
feature young child sit dining table enjoy piece cake child eat cake fork knife nearby table surround chair chair left table right addition cake cup place table possibly drink accompany dessert scene capture moment joy indulgence child

Annoyance, Anticipation, Aversion, Disquietment, Engagement, Excitement, Happiness, Sadness

feature sit desk wear suit tie appear state distress discomfort hand forehead surround item desk include cup cell phone book pen cup place desk near left near right scene suggest work deal challenging situation

Annoyance, Anticipation, Aversion, Disquietment, Engagement, Excitement, Happiness, Pain, Sadness

Fig. 4. The GIN pipeline exemplified for the 4 pictures. The captions are post-processing and the labels are the ones predicted by the model



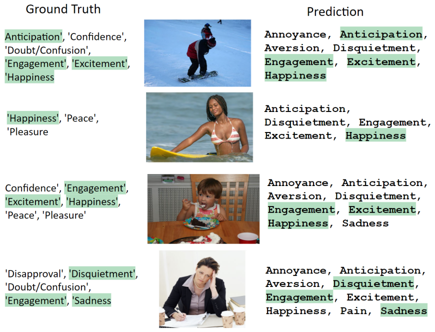
The image features a young child sitting at a dining table, enjoying a piece of cake. The child is using a fork to eat the cake, which is placed on a plate in front of them. The table is surrounded by chairs, with one chair on the left side of the table and another on the right side. \n\nIn addition to the cake, there are two cups on the table, one near the right edge and the other closer to the center. A knife can also be seen on the table, likely used for cutting the cake. The scene captures a moment of joy and indulgence for the child as they savor their treat.

feature young child sit dining table enjoy piece cake child eat cake fork knife nearby table surround chair chair left table right addition cake cup place table possibly drink accompany dessert scene capture moment joy indulgence child

The image features a woman sitting at a desk, wearing a suit and tie. She appears to be in a professional setting, possibly working or studying. She is holding a pen in her hand, possibly taking notes or working on a task. \n\nThere are several cups placed on the desk, with one cup located near the woman and the others scattered around the desk. A cell phone can be seen on the desk as well, indicating that the woman might be using it for communication or organization purposes. The scene suggests a focused and productive environment.

feature sit desk wear suit tie appear state distress discomfort hand forehead surround item desk include cup cell phone book pen cup place desk near left near right scene suggest work deal challenging situation

Fig. 5. Two of the captions before and after processing



| Ground Truth | | Prediction |
|---|---|---|
| Anticipation', 'Confidence', 'Doubt/Confusion', 'Engagement', 'Excitement', 'Happiness | | Annoyance, Anticipation, Aversion, Disquietment, Engagement, Excitement, Happiness |
| 'Happiness', 'Peace', 'Pleasure | | Anticipation, Disquietment, Engagement, Excitement, Happiness |
| Confidence', 'Engagement', 'Excitement', 'Happiness', 'Peace', 'Pleasure' | | Annoyance, Anticipation, Aversion, Disquietment, Engagement, Excitement, Happiness, Sadness |
| 'Disapproval', 'Disquietment', 'Doubt/Confusion', 'Engagement', 'Sadness | | Annoyance, Anticipation, Aversion, Disquietment, Engagement, Excitement, Happiness, Pain, Sadness |

Fig. 6. The predictions made by GIN and the annotations of the images

### 3.2 Experimental Setup

For this investigation, 2444 images from the test split of the EMOTIC dataset were used. This dataset was selected due to its wide range of images depicting people in real-world environments, making it particularly suitable for evaluating both the zero-shot emotion recognition performance of LLaVA and the performance of the Graph Isomorphism Network (GIN). Its rich contextual diversity allows LLaVA to generate detailed descriptions, which are crucial for the graph-based reasoning of the GIN. Additionally, the GIN developed by Willams deLima Costa was originally trained and evaluated on EMOTIC, making it the most appropriate choice for direct comparison.

The 2444 chosen images were selected for this investigation because they represent the subset of "single-subject images” from the test split of the dataset. The term “single-subject images” used here refers to images in which there is only one list of emotions annotations per image. Since in the EMOTIC database each picture is annotated with a list of emotions per individual present in the image, the pictures used in the study were filtered such that a single emotions annotation list is present per image. As a result, the majority of these images depict a single individual, although occasionally some may include multiple people, either with one person clearly in focus or with all individuals sharing the same annotated emotion list.

### 3.3 Implementation Details

When predicting multiple emotions against the EMOTIC annotations, the performance of LLaVA was evaluated using the micro-averaged F1 score. This evaluation metric balances the number of correct predictions with the number of wrong and missing predictions. This result was then compared with the performance (micro-averaged F1 score) of the Graph Isomorphism Network.

The Graph Isomorphism Network was configured with a 50-dimensional input feature vector, a 40-dimensional hidden layer, and a 26-unit output layer corresponding to the emotion categories, and its pre-trained weights were loaded from the GitHub repository of the paper. For evaluation, two primary metrics were used: the mean Average Precision (mAP) and the micro-averaged F1 score, both computed using standard functions from the *sklearn.metrics* module. Since the computation of the F1 score requires binary predictions rather than raw confidence scores, thresholding was necessary. The GIN model outputs raw logits, where large positive values indicate high confidence in the presence of an emotion and large negative values suggest high confidence in its absence. These logits must therefore be binarized by applying a threshold to enable evaluation against the ground truth labels using the F1 score. Two binarization strategies were explored:

1. Zero Thresholding: In the first approach, predictions were binarized using a threshold of 0. Any logit above 0 was mapped to 1 (positive prediction), and anything below or equal to 0 was mapped to 0 (negative prediction).
2. High Confidence Thresholding: A more conservative threshold of 500 was used after experimenting with various values and inspecting raw logit magnitudes. This threshold struck a balance between filtering out uncertain predictions and retaining a meaningful number of categorical outputs.

### 3.4 Results

Table 2 shows the performance of both LLaVA and Graph Isomorphism Network in the task of multi-label emotion prediction on the EMOTIC dataset. The result of the GIN using both binarization methods mentioned in the *Implementation Details* section is shown. Although the GIN was evaluated using five different captions, only the highest obtained score is presented in this table. The performance of LLaVA is measured using the F1 score, whereas the performance of the GIN is measured using both the F1 score and the mAP. The reason for this is that the mAP metric requires confidence scores to be computed, which are not provided in the predictions of LLaVA. The table also shows the performance of the GIN announced in the "High-level context representation for emotion recognition in images" paper [5].

Table 3 presents the performance of the GIN model using five different captions, evaluated under both binarization strategies. The results highlight the influence of variations in the input captions on the performance of the model.

Table 2. LLaVA vs GIN performance in multi-emotion prediction on the EMOTIC dataset

| **Model** | **mAP** | **Micro-averaged F1 Score** |
|---|---|---|
| Zero-shot LLaVA | – | 0.4065 |
| GIN, BinS1 | 0.2012 | 0.3841 |
| GIN, BinS2 | 0.2012 | 0.8107 |
| ExpansionNet-v2 + GIN | 0.3002 | – |

*Note:* BinS1 refers to the Binarization Strategy 1 and BinS2 refers to Binarization Strategy 2

Table 3. GIN performance with five different input captions and two binarization strategies

| **Model** | **F1 using BinS1** | **F1 using BinS2** | **mAP** |
|---|---|---|---|
| C0 + GIN | 0.3841 | 0.8107 | 0.2012 |
| C1 + GIN | 0.3826 | 0.7205 | 0.1989 |
| C2 + GIN | 0.3805 | 0.7366 | 0.2012 |
| C3 + GIN | 0.3780 | 0.7349 | 0.1971 |
| C4 + GIN | 0.3787 | 0.7016 | 0.1921 |

*Note:* BinS1 refers to the Binarization Strategy 1 and BinS2 refers to Binarization Strategy 2
C0,C1..C4 represent the different captions used

## 4 SENTIMENT RECOGNITION IN IMAGES: LLAVA ON THE FINDINGEMO DATASET

### 4.1 Methodology

This investigation used the FindingEmo dataset and followed the methodology outlined below.

For each image, the zero-shot LLaVA model was prompted to identify the representative sentiment from a set of 24 possible options, corresponding to the sentiment labels defined in the FindingEmo dataset.

Accuracy was used as the evaluation metric, and the results were compared against a random baseline, as well as the reported state of the art on the FindingEmo dataset, in order to contextualize the performance of the model. An illustration of this process is provided in Figure 7.

### 4.2 Experimental Setup

For this investigation, a total of 2,943 images from the FindingEmo dataset were used. This dataset was chosen because it is one of the few publicly available resources that provides annotations for the overall sentiment of an image, rather than assigning individual emotion labels to each person depicted within it.

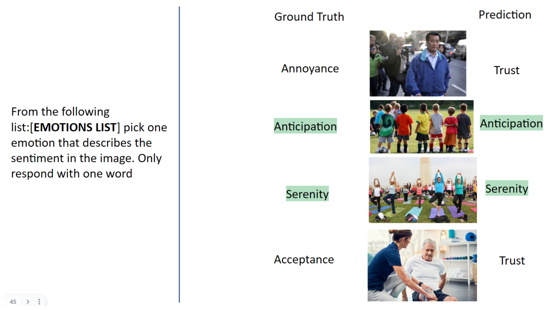

From the following list:**[EMOTIONS LIST]** pick one emotion that describes the sentiment in the image. Only respond with one word

Fig. 7. The predictions generated by LLaVA and labels from the FindingEmo dataset, as well as the prompt used

### 4.3 Implementation Details

For evaluating the predictions of the model for the single-label images in the FindingEmo dataset, accuracy was measured. Two evaluation approaches were employed: one based on the full set of 24 annotated emotions (Emo24) and another using a reduced set of 8 grouped emotions (Emo8).

For Emo24, model performance was computed by comparing predictions against the ground truth across all 24 possible emotions per image. In contrast, for Emo8, both the model predictions and the annotated labels were mapped to one of the 8 broader emotional categories defined by *Plutchik's Wheel of Emotions* [14]. This mapping reduced the level of granularity required for a prediction to be considered correct.

### 4.4 Results

For Emo24, the performance of LLaVA on the sentiment detection task is illustrated in Table 4. To contextualize the results of LLaVA, its accuracy was compared against a random baseline. Since no standardized or peer-reviewed performance benchmarks for Emo24 were found in the original FindingEmo paper or subsequent literature, this random baseline serves as a practical reference for comparison.

Table 4. Accuracy of single emotion prediction on the FindingEmo dataset - Emo24 and accuracy of a random baseline

| **Model** | **FindingEmo dataset (Emo24)** |
|---|---|
| **Zero-shot LLaVA** | 19.78% |
| **Random Baseline*** | 4.17% |

*Let $P_{correct\ guess} = \frac{1}{24} = 0.0417$.
*Expected correct guesses:* $E = Total_{Nb} \times P_{correct\ guess} = 2943 \times 0.0417 = 122.625$.
*Expected accuracy:* $Acc = \frac{Correct\ Guesses}{Total_{Nb}} = \frac{122.625}{2943} = 0.04166 = 4.17\%$.

For Emo8, the performance of LLaVA is shown in Table 5. The state-of-the-art performance obtained by the CLIP model as reported in the original FindingEmo paper is also shown in the table, for comparison [14].

Table 5. Accuracy of single-emotion prediction on the FindingEmo dataset (Emo8), compared with the state-of-the-art CLIP model performance as reported in the original FindingEmo paper.

| Model | FindingEmo dataset (Emo8) |
|---|---|
| **Zero-shot LLaVA** | 36.77% |
| **Baseline CLIP** | 43% |

## 5 DISCUSSION

In this section, answers to the posed research questions will be provided, as well as the limitations of the work and possible future improvements.

Firsly, the results of the initial LLaVA experiment will be disscused. LLaVA was evaluated on single emotion prediction within multi-labeled images from the EMOTIC dataset (see Table 1). In this case, the observed accuracy was considerably high at 61%. However, when benchmarked against a random baseline that achieved 52%, the margin of improvement appears less significant. This is expected, as predicting one relevant emotion out of several annotated ones is an intuitively not very challenging task. Although the margin of improvement is modest, the zero-shot performance of LLaVA remains notable, demonstrating its potential for emotion recognition tasks.

### 5.1 Answers to the Research Questions

#### Answers to Research Question 1

To compare the performance of LLaVA with that of the Graph Isomorphism Network (GIN) in the task of emotion recognition, the results presented in Table 2 are analyzed. The table shows that GIN achieved a much higher F1 score than LLaVA when using the second binarization strategy: High Confidence Thresholding. However, under the first binarization strategy, the standalone performance of LLaVA was higher than that of GIN. These results indicate that, while LLaVA performs moderately well in a zero-shot setting, the GIN can significantly outperform LLaVA in categorical emotion prediction if a more restrictive binarization strategy is applied.

It is also worth noting that, as shown in Table 2, the performance of the GIN when using LLaVA-generated captions is lower than the results reported in the original study. This discrepancy is likely due to the mismatch between the vocabulary used for training the model (derived from ExpansionNet captions), and the vocabulary present in the LLaVA captions, used for testing the model.

Additionally, from Table 3 it is also visible that with zero thresholding (Binarization Strategy 1), the model greatly overpredicts and achieves a lower F1 score due to many false positives, while high confidence thresholding (Binarization Strategy 2) substantially improves precision by filtering out low-confidence predictions, ultimately resulting in a significantly higher and more reliable micro-averaged F1 score.

Furthermore, using slightly different captions as input to the GIN model results in only minor variations in F1 score and mAP, as it can be observed in Table 3, for both of the binarization strategies.

#### Answers to Research Question 2

Using Table 4 to analyze the Emo24 results, the performance of the model in single-label emotion detection may initially seem low, with an accuracy of approximately 20%. However, this outcome reflects the inherent difficulty of the task, which requires the model to identify the single correct emotion out of 24 possible categories. When compared to a random baseline, however, the improvement is substantial, particularly given that LLaVA was evaluated in a zero-shot setting without any task-specific fine-tuning.

Moreover, as the FindingEmo dataset was only released in mid-2024, there is a lack of available studies reporting performance on the Emo24 task. This is also likely due to the considerable challenge posed by the single-label prediction requirement. This difficulty is also acknowledged in the original FindingEmo paper [14].

Given these comparisons, the 20% accuracy obtained in the more challenging single-labeled setting is actually a promising result, and an important baseline for future work, reflecting the potential of LLaVA in complex zero-shot emotion recognition tasks.

From Table 5, it is notable that LLaVA performed comparably to the state-of-the-art CLIP model on the FindingEmo dataset, in the Emo8 context. The F1 score is significantly higher than in the Emo24 setting, which is expected given that the task involves only 8 possible emotion categories, making it inherently simpler. These results further demonstrate that, in the task of sentiment recognition, zero-shot LLaVA shows good overall performance.

#### Answer to Main Research Question

The analysis of the performance of LLaVA across multiple tasks and evaluation settings demonstrates that LLaVA has promising capabilities for emotion recognition from images, particularly in zero-shot scenarios. While its performance in the Emo24 task (single-label classification with 24 categories) is modest, it still significantly surpasses the random baseline, highlighting its potential despite the complexity of the task. Furthermore, in the Emo8 scenario, LLaVA performs competitively with the state-of-the-art CLIP model.

However, on the EMOTIC dataset, results also show that the GIN model can outperform LLaVA in multi-label emotion prediction. This suggests that, while LLaVA is a capable generalist model, fine-tuned or context-aware models such as the GIN, may still outperform LLaVA in emotion prediction tasks.

Overall, LLaVA proves to be a strong baseline for zero-shot emotion recognition and demonstrates that vision-language models can capture emotional cues from visual content even without task-specific training.

### 5.2 Limitations and Observations

This section outlines the limitations encountered during the research and highlights important observations made throughout the experimental process.

#### Limitations

**Limitation 1:** The GIN model used for the experiments was pre-trained on captions generated by the ExpansionNet-v2 model [8], but tested on captions generated by the vision-language model LLaVA [12] in this research.

This introduces a mismatch, as the vocabulary and linguistic patterns in the training corpus differ significantly from those in the LLaVA-generated captions used during testing. The LLaVA captions tend to be more expressive and often include emotion-related words and references to facial expressions—elements that are absent from ExpansionNet-generated captions.

**Limitation 2:** Some of the labels in the annotations were wrong. An example of this is illustrated in Appendix A Figure 8. Such inaccuracies inevitably affect the evaluation process, leading to a seemingly poor model performance, even when the prediction of the model is actually reasonable or correct. In these cases, the discrepancy arises not from a model error but from an unreliable or questionable ground truth label.

**Limitation 3:** Emotion classification is subjective, therefore an objective measurement for this task is not appropriate.

Emotion detection and the evaluation of related models are inherently challenging and subject to bias, primarily because the emotional content of an image can be interpreted differently by different annotators [13]. Consequently, benchmarking model predictions against annotated "ground truth" labels is not always a fair assessment. In many cases, the prediction of the model may be entirely reasonable, or even more accurate, yet still considered incorrect simply because it does not match the annotation.

To further investigate this issue, a small-scale survey was conducted with 27 participants using 10 images from the FindingEmo dataset. These images were manually selected based on cases where the predictions of LLaVA differed from the annotated labels but were actually more appropriate than the ground truth annotations.

For each image, participants were asked to select the emotion that best describes the sentiment portrayed. Two options were provided: the original ground truth label and the prediction of LLaVA. The results show that for 8 out of the 10 images, a clear majority of participants chose the prediction of the model over the ground truth. Remarkably, for one image, 100% of participants agreed with the prediction of the model. A summary of these results is presented in the Appendix A, Table 6. From these results, it is clear that, although counted as incorrect, it is possible that the predictions of the model are, in reality, appropriate and can be perceived as more representative than the ground truth annotations. This supports the claim that emotion recognition is subjective, and objective evaluation is often not appropriate for the task.

## Observations

An interesting area explored in parallel with the main research is the prompt sensitivity of vision-language models. To investigate how variations in prompt formulations influence the outputs and performance of the model, several supplementary experiments were conducted throughout this study. These experiments were guided by targeted subquestions aimed at examining specific aspects of prompt influence:

*How do different prompt formulations influence the image captions generated by the model?*

To address this question, several distinct prompt formulations were tested, each phrased differently but aiming to produce a descriptive caption of the same image. The intent was to observe how linguistic variation in the prompt affects the detail, focus, or emotional tone of the response of the model. An example is provided in Appendix A, in Figure 9, showcasing an image alongside the various prompts used and the corresponding captions generated by the model.

From the experiments, it appears that using different prompts for image captioning with LLaVA does not substantially affect the generated captions. As shown in the example in Appendix A, in Figure 9, only minor variations are present, which can be considered insignificant. This finding is further supported by the insignificant variation in performance of the Graph Isomorphism Network (GIN) when using captions derived from different prompts, shown in Table 3. It is reasonable to assume that, if the produced captions had significant differences, the performance of the GIN would also vary significantly.

*Does changing the order of the possible emotions influence the predictions of the model?*

To examine this, the experimental setup described for the FindingEmo investigation was reused. This included the 24 emotion labels from the FindingEmo dataset and the same set of test images. The model was executed twice, each time with the list of candidate emotions arranged in a different order. The resulting predictions were again evaluated using accuracy to determine whether the order of the labels affected the performance of the model.

After shuffling the order of the possible emotions from the prompt, the accuracy of the model decreased from an initial 19.78% to 17.02%. This decrease is significant and it suggests that even seemingly minor changes, such as the ordering of options, can meaningfully affect the performance of the model.

*How well can the model understand numeric boundaries indicated in complex prompts?*

While addressing the first research question, various prompt formulations were experimented with before applying them to the 2,444 images from the EMOTIC dataset. The final prompt used was:

*"Look at the person in the image and choose all emotions they are feeling from this list (they may feel more than one): [EMOTION LIST]. Only list the relevant emotions, maximum 8. Do not explain."*

This specific formulation was reached after testing several alternatives with different numeric constraints.

Initially, no numerical limitation was included in the prompt, which led the model to produce predictions that either included all 26 emotions or only a single one per image, both of which were undesirable. To address this, numerical boundaries were explicitly introduced. It was observed that the model handled small numerical limits (e.g., fewer than 8) well, consistently generating a number of emotions below or equal to the specified maximum. However, for values above 8, the model typically reverted to the same extremes: predicting either all 26 or only one emotion. Interestingly, when the limit was set to exactly 8, the model sometimes predicted up to 10 emotions, slightly exceeding the boundary but remaining within the limit of the EMOTIC annotations, where no image has more than 10 labeled emotions. This made the value of 8 a convenient and effective choice for evaluation purposes.

These findings indicate that the model demonstrates limited and inconsistent adherence to numeric constraints in complex prompts.

While it can follow lower numerical boundaries with relative accuracy, it often fails to enforce higher ones, suggesting that it only partially understands or applies such instructions in a reliable way.

*General Conclusion for Prompt Sensitivity*

These findings indicate that prompt formulation plays an important role in nuanced, multi-step tasks, or tasks that require complex reasoning, such as selecting relevant emotions from an extensive predefined list. In such cases, variations in phrasing can lead to substantially different outputs. Conversely, for simpler tasks like generating descriptive captions, the influence of prompt wording is minimal.

## 6 FUTURE WORK

While this research has provided valuable insights into the capabilities of LLaVA for the task of emotion recognition, several areas remain open for further exploration and improvement. The following directions outline potential areas of research for future work:

**Training the GIN on LLaVA captions**: As mentioned in the *Limitations* section, the GIN model used for the experiments in this research was pre-trained on captions generated by the ExpansionNet-v2 model [8], but tested on captions generated by the vision-language model LLaVA [12]. As this is problematic because of the vocabulary differences, it would be a good approach for future research to train the GIN on LLaVA-generated captions, and then evaluate its performance, comparing it to the reported performance of the GIN in this paper as well as the original one. Furthermore, LLaVA could also be fine-tuned on the EMOTIC and FindingEmo datasets, and performance could be compared to the one reported in the present research.

**Emotion Order Sensitivity**: Given the experimental findings that the order of emotions in the candidate list appears to influence model performance, this aspect should be investigated further. One promising direction is to systematically explore how the position of specific emotions in the list affects their likelihood of being predicted. For example, it would be valuable to assess whether placing the most frequently predicted emotions at the end of the list leads to a decrease in their selection. Such experiments could help reveal whether the model exhibits positional biases and whether performance can be optimized by reordering the emotion labels.

**Bias and Fairness Analysis**: Given the inherent subjectivity involved in emotion annotation, future work should also examine whether the predictions of the model exhibit biases related to demographic, cultural, or contextual factors. Identifying and addressing such biases is important for ensuring fairness in real-world applications of emotion recognition systems [13].

## 7 CONCLUSION

This study explored the performance of the zero-shot, state-of-the-art vision-language model LLaVA in the task of emotion recognition. The model was evaluated on two representative datasets: EMOTIC and FindingEmo, and the results were compared to the performance of the pre-trained Graph Isomorphism Network model developed by Willams de Lima Costa [5] on the EMOTIC dataset, as well as the state-of-the-art CLIP model on the FindingEmo dataset, as reported in the original paper [14]. The findings show that zero-shot LLaVA performs well in the task of emotion recognition, but it is still outperformed by a model pre-trained for this task, such as the GIN. The research also outlines the difficulty of the emotion recognition task, highlighting the subjectivity of labeling, which can sometimes be inaccurate or inappropriate. Additionally, it explores the sensitivity of the model to slight changes in prompts and highlights the limitations of evaluating a GIN on a different vocabulary than the one it had been trained on.

## REFERENCES

[1] Andoni Beristain and Manuel Graña. 2009. Emotion recognition based on the analysis of facial expressions: A survey. *New Mathematics and Natural Computation* 5, 02 (2009), 513–534. https://doi.org/10.1142/S1793005709001453

[2] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. 2024. An Introduction to Vision-Language Modeling. arXiv:2405.17247 [cs.LG] https://arxiv.org/abs/2405.17247

[3] Erik Cambria, Soujanya Poria, Ivan Vulić, Amir Hussain, and Björn Schuller. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. *Information Fusion* 63 (2020), 143–153.

[4] Willams Costa, David Macêdo, Cleber Zanchettin, Estefanía Talavera, Lucas Silva Figueiredo, Veronica Teichrieb, and João Marcelo Teixeira. 2025. A Fast Multiple Cue Fusing Approach for Human Emotion Recognition. https://doi.org/10.2139/ssrn.5205383 ssrn:5205383 SSRN preprint.

[5] Willams de Lima Costa, Estefania Talavera Martinez, Lucas Silva Figueiredo, and Veronica Teichrieb. 2023. High-Level Context Representation for Emotion Recognition in Images. arXiv:2305.03500 [cs.CV] https://arxiv.org/abs/2305.03500

[6] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[7] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual Emotion Recognition using Large Vision Language Models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4769–4776. https://doi.org/10.1109/iros58592.2024.10802538

[8] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2023. Exploiting Multiple Sequence Lengths in Fast End to End Training for Image Captioning. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2173–2182. https://doi.org/10.1109/bigdata59044.2023.10386812

[9] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context Based Emotion Recognition using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. https://doi.org/10.1109/tpami.2019.2916866

[10] Jiangjian Li, Tao Sun, Liang Wang, Tieniu Tan, and Zhaoxiang Zhang. 2022. CAGNet: A Context-Aware Graph Neural Network for Detecting Social Relationships in Videos. *arXiv preprint arXiv:2211.07947* (2022). arXiv:2211.07947 [cs.CV] https://arxiv.org/abs/2211.07947

[11] Fatiha Limami, Boutaina Hdioud, and Rachid Oulad Haj Thami. 2024. Contextual emotion detection in images using deep learning. *Frontiers in Artificial Intelligence* Volume 7 - 2024 (2024). https://doi.org/10.3389/frai.2024.1386753

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] https://arxiv.org/abs/2304.08485

[13] Oscar Martinez-Miwa and Mauricio Castelán. 2023. On reliability of annotations in contextual emotion imagery. *Scientific Data* 10, 1 (2023), 525. https://doi.org/10.1038/s41597-023-02435-1

[14] Laurent Mertens, Elahe' Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. 2024. FindingEmo: An Image Dataset for Emotion Recognition in the Wild. arXiv:2402.01355 [cs.CV] https://arxiv.org/abs/2402.01355

[15] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-Aware Multimodal Emotion Recognition using Frege's Principle. arXiv:2003.06692 [cs.CV] https://arxiv.org/abs/2003.06692

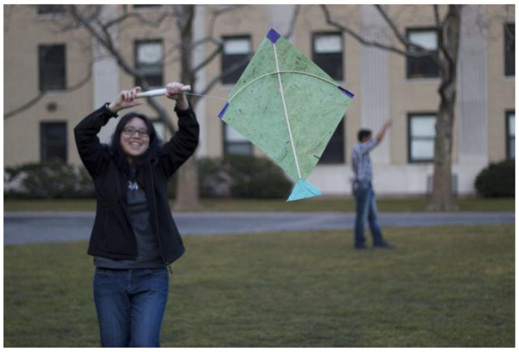
[16] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K. Roy-Chowdhury. 2024. Unbiased Scene Graph Generation in Videos.

[17] George Ordoumpozanis and Grigorios A. Papakostas. 2025. Reviewing 6D Pose Estimation: Model Strengths, Limitations, and Application Fields. *Applied Sciences* (2025).

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

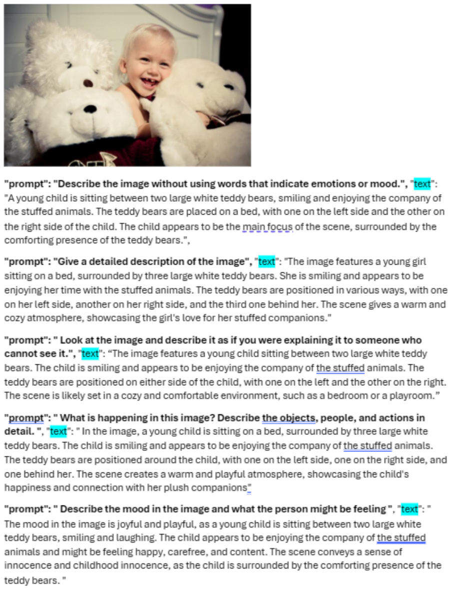
Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

## A ADDITIONAL FIGURES AND TABLES



Emotions: ['Confidence', 'Engagement', 'Excitement', 'Happiness', 'Pleasure', 'Suffering']

Fig. 8. An image from the EMOTIC dataset where the red annotated emotion is clearly not suited for the image



**"prompt": "Describe the image without using words that indicate emotions or mood.", "text":** "A young child is sitting between two large white teddy bears, smiling and enjoying the company of the stuffed animals. The teddy bears are placed on a bed, with one on the left side and the other on the right side of the child. The child appears to be the main focus of the scene, surrounded by the comforting presence of the teddy bears.",

**"prompt": "Give a detailed description of the image", "text":** "The image features a young girl sitting on a bed, surrounded by three large white teddy bears. She is smiling and appears to be enjoying her time with the stuffed animals. The teddy bears are positioned in various ways, with one on her left side, another on her right side, and the third one behind her. The scene gives a warm and cozy atmosphere, showcasing the girl's love for her stuffed companions."

**"prompt": " Look at the image and describe it as if you were explaining it to someone who cannot see it.", "text":** "The image features a young child sitting between two large white teddy bears. The child is smiling and appears to be enjoying the company of the stuffed animals. The teddy bears are positioned on either side of the child, with one on the left and the other on the right. The scene is likely set in a cozy and comfortable environment, such as a bedroom or a playroom."

**"prompt": " What is happening in this image? Describe the objects, people, and actions in detail. ", "text":** " In the image, a young child is sitting on a bed, surrounded by three large white teddy bears. The child is smiling and appears to be enjoying the company of the stuffed animals. The teddy bears are positioned around the child, with one on the left side, one on the right side, and one behind her. The scene creates a warm and playful atmosphere, showcasing the child's happiness and connection with her plush companions"

**"prompt": " Describe the mood in the image and what the person might be feeling ", "text":** " The mood in the image is joyful and playful, as a young child is sitting between two large white teddy bears, smiling and laughing. The child appears to be enjoying the company of the stuffed animals and might be feeling happy, carefree, and content. The scene conveys a sense of innocence and childhood innocence, as the child is surrounded by the comforting presence of the teddy bears. "

Fig. 9. An image from the EMOTIC dataset and various prompts and the resulting captions

Table 6. Survey results, where the green color indicates the bigger percentage

| | Label, Percentage | Prediction, Percentage |
|---|---|---|
| Image 1 | Interest, 74.1% | Joy, 25.9% |
| Image 2 | Acceptance, 33.3% | Trust, 66.7% |
| Image 3 | Interest, 25.9% | Anger, 74.1% |
| Image 4 | Ecstasy, 7.4% | Joy, 92.6% |
| Image 5 | Acceptance, 22.2% | Joy, 77.8% |
| Image 6 | Serenity, 33.3% | Joy, 66.7% |
| Image 7 | Sadness, 59.3% | Anticipation, 40.7% |
| Image 8 | Anticipation, 37% | Fear, 63% |
| Image 9 | Joy, 0% | Trust, 100% |
| Image 10 | Annoyance, 18.5% | Vigilance, 81.5% |