


# EmoViT: Revolutionizing Emotion Insights with Visual Instruction Tuning

## Supplementary Material

**Our Emotion Instruction Data Sample**



**Context type 1: Caption:**  
A woman in a red dress singing into a microphone.

**Context type 2: Attribute:**  
 Emotion: excitement    Human action: singing  
 Brightness: 0.6        Object: ['Fashion accessory']  
 Colorfulness: 0.9      Facial expression: excited

---

**Categorical**

**Question:** Please select the emotion closest to the image from the following options: amusement, anger, awe, contentment, disgust, excitement, fear and sadness.  
**Answer:** Predict emotion: excitement.

**Conversation    Basic Interaction**

**Question:** What is the main action the woman in the image is performing?  
**Answer:** The main action that the woman in the image is performing is singing into a microphone.

**Conversation    Advanced Interaction**

**Question:** Can you discuss the possible occasion or event happening in the image based on the woman's attire and action?  
**Answer:** Given the woman's attire and action, it can be inferred that she is likely participating in a music-related event. Her elegant red dress and fashion accessory suggest a festive or formal gathering, perhaps a concert, party, or a similar event.

**Reasoning**

**Question:** Given the woman's attire, activity, and the palpable excitement, what could be the nature of the event she's participating in and what factors are contributing to the atmosphere of excitement?  
**Answer:** Live music performance naturally tends to generate excitement due to the direct interaction between the performer and audience. The audience's reception and response to the music may be driving the performer's excitement.

Figure 10. The sample of our generated visual emotion instruction data.

### 1. More Emotion Visual Instruction Data Sample


Additional samples from our Emotion Visual Instruction Data collection are presented in Figures 10 and 11. Upon acceptance, the complete dataset will be made available on our project webpage.

## 2. Implemental Details

### 2.1. Our Experiment Settings

*Held-out vs supervised learning.* We adopt the terminology *held-in* and *held-out* as defined in the work of InstructBLIP [1]. For the held-in, we utilize the training subset of the EmoSet dataset for Emotion Visual Instruction Tuning, with its corresponding test subset serving the purpose of held-in evaluation. The outcomes of this evaluation are depicted in Fig. 1 of the main manuscript.

**Our Emotion Instruction Data Sample**



**Context type 1: Caption:**  
A young boy with a red turban on his head leaning on a wooden fence.

**Context type 2: Attribute:**  
 Emotion: sadness        Human action: stretching arm  
 Brightness: 0.6        Object: ['Tree']  
 Colorfulness: 0.7      Facial expression: sad

---

**Categorical**

**Question:** Please select the emotion closest to the image from the following options: amusement, anger, awe, contentment, disgust, excitement, fear and sadness.  
**Answer:** Predict emotion: sadness.

**Conversation    Basic Interaction**

**Question:** Are there any other people or animals in the image apart from the young boy?  
**Answer:** No, there are no other people or animals in the image. The sole human figure present is the young boy.

**Conversation    Advanced Interaction**

**Question:** Can you infer anything about the boy's environment and way of life from the image?  
**Answer:** His location near a wooden fence might imply a rural or village setting. The presence of a tree in the background might imply he is outdoors in a natural environment, perhaps a wooded or countryside area.

**Reasoning**

**Question:** Considering the incident of the boy with a red turban expressing sadness while leaning on a wooden fence, can you speculate possible reasons behind his melancholy and how it might relate to the surrounding environment?  
**Answer:** The young boy may be experiencing a sense of loneliness or distress, represented by his sad expression and the act of leaning against the fence as if seeking comfort or support.

Figure 11. The sample of our generated visual emotion instruction data.

In our held-out evaluation, we focus on determining how instruction tuning bolsters the model's ability to transfer learning to new and unseen data. It's crucial to highlight that our methodology sets a distinct path from InstructBLIP's framework. Our dataset is specifically curated with emotion-centric content, presenting unique categories such as *cheerfulness* and *enthralment* found in WEBEmo, which are not typically included in other datasets. Conversely, common emotional categories like *anger* and *fear* are shared with other collections, such as FI and Emotion6. This distinctive mix in our dataset implies that our held-out evaluation operates on a *cross-domain* level, examining the model's ability to interpret and adapt to diverse emotional contexts not strictly confined to zero-shot scenarios.

### 2.2. System Prompt

The system prompt inputted into ChatGPT for the purpose of gathering instruction-based data is presented below.

You are an AI visual assistant, and you are seeing a single image. What you see are provided with one caption and some emotion related attributes, describing the same image you are looking at. Answer all questions as you are seeing the image. The range of brightness is from 0 (darkest) to 1 (brightest), and the range of colorfulness is from 0 (black-and-white) to 1 (the most colorful).

Design two questions for a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, object actions, relationship among objects, etc. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently. Please answer with the format *Question: Answer:*

Also include one complex question that is relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

### 2.3. Details of the Q-Former

Similar to the approach in InstructBLIP, Q-Former is a lightweight transformer architecture that utilizes a collection of trainable query vectors to distill visual features from a static image encoder. The Q-Former acts as the trainable module to bridge the gap between a frozen image encoder and a frozen LLM. Its role is to curate and present the most pertinent visual information, thereby enabling the LLM to generate the targeted textual output efficiently. Following the default setting, in our experimental setup, we employ 32 distinct queries, each with a dimensionality of 768.

### 2.4. Sensitivity Formula

As mentioned in Sec.4.3.2 in the main paper, we employ the *Sensitivity* evaluation metric, as introduced by [2], to assess the model’s fidelity in generating uniform outcomes irrespective of instructional nuances. Specifically, for each task  $t \in T$ , given its associated instances with task instruc-

tions:  $D^t = \{(I_j^t, x_j^t, y_j^t) \in T \times X^t \times Y^t\}_{j=1}^N$ , sensitivity is defined as:

$$\mathbf{E}_{t \in T} \left[ \frac{\sigma_{i \in I^t} [\mathbb{E}_{(x,y) \in D^t} [\mathcal{L}(f_\theta(i, x), y)]]}{\mu_{i \in I^t} [\mathbb{E}_{(x,y) \in D^t} [\mathcal{L}(f_\theta(i, x), y)]]} \right] \quad (1)$$

where  $\mathcal{L}$  denotes the evaluation metric, *i.e.*, emotion classification accuracy,  $f_\theta(\cdot)$  represents the Visual Instruction Tuning model. The standard deviation and mean of the model’s performance across all instructions are denoted by  $\sigma_{i \in I^t}[\cdot]$  and  $\mu_{i \in I^t}[\cdot]$ , respectively.

## 3. Ablation Study of LLM Model Size

In our attempts with the EmoVIT architecture’s LLM, we explored the use of models of varying sizes (as shown in Tab. 5). The results indicated that the smaller model, Vicuna7B, outperformed its larger counterparts. This may be attributed to the limited training data available for our task, which potentially underutilizes the capabilities of larger models. Consequently, we anticipate that an increase in training data in the future could enhance the effectiveness of Emotion Visual Instruction Tuning.

Vicuna-7B	Vicuna-13B	FlanT5XL
83.36	82.21	80.98

Table 5. Ablation study of different LLM model size. Accuracy (%) on EmoSet test set.

## 4. GPT-4 vs GPT-4 Turbo

We conducted a comparative analysis of conversational datasets derived from GPT-4 (the model name is *gpt-4* in the API) against the recently released GPT-4 Turbo (the model name is *gpt-4-1106-preview* in the API). The comparative metrics yielded negligible differences between the two models (83.36% vs 82.96% on EmoSet test set).

## 5. Adding In-context Samples in Held-out Evaluation

Recent LLMs are capable of in-context learning when provided with a limited number of examples in a few-shot manner. In this work, we have also embarked on such an exploration. For instance, Tab. 6 presents the in-context samples utilized within the EmotionROI dataset. During our held-out evaluation, we incorporated three in-context samples for each category, consisting of a caption paired with its corresponding emotion class. Nevertheless, in our experimental observations, we did not witness any enhancement in performance attributable to furnishing the LLM with these in-context examples. Consequently, our finalized methodology did not incorporate in-context samples during the held-out evaluation phase.

Description	Emotion
Unleashed Fury: A portrait of raw, unfiltered anger etched on the subject’s face.	Anger
Volcanic Eruption in Human Form: A Portrait of Unrestrained Fury.	Anger
An explosive portrait of raw fury, where every clenched jaw and furrowed brow tells a tale of unchecked anger.	Anger
Face contorted in a grimace of pure disgust, as if they just tasted a year-old lemon.	Disgust
Caught in the throes of revulsion, a face grimaces as if it just tasted the world’s sourest lemon.	Disgust
Picture Perfect: A Masterclass in the Art of Disgust Expression	Disgust
A chilling moment of pure terror, etched in every detail.	Fear
A chilling moment of pure terror etched on the face, a stark embodiment of fear.	Fear
someone with a wide smile, a group	Joy
Overflowing with joy, like a puppy at a park!	Joy
A poignant portrait of sorrow, where teardrops are the silent language of grief.	Sadness
An evocative portrayal of sorrow, with shadows seemingly swallowing the light, reflecting the heavy weight of sadness.	Sadness
An abstract portrayal of solitude, where the vivid hues of melancholy paint a poignant picture of sadness.	Sadness
Caught in a moment of pure astonishment, eyes wide and mouth agape.	Surprise
Caught in the headlights of astonishment: a jaw-dropping moment of surprise!	Surprise
Caught in the Act! A person’s wide-eyed gasp of sheer surprise.	Surprise

Table 6. Illustrative Examples of Emotion Descriptors in Visual Data

## 6. Limitation and future work

Due to the reliance on the GPT-API and cost considerations, our held-in pretraining phase utilized less than 50% of the EmoSet dataset. Despite outperforming other methods, we recognize the potential for significant improvements in future work by expanding the data scale. We anticipate that advancements in visual emotion understanding will parallel increases in both data and model scale.

## References

- [1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.
- [2] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2022.