

# Affective Visual Dialog: A Large-Scale Benchmark for Emotional Reasoning Based on Visually Grounded Conversations

Kilichbek Haydarov<sup>1</sup>, Xiaoqian Shen<sup>1</sup>, Avinash Madasu<sup>1</sup>, Mahmoud Salem<sup>1</sup>, Li-Jia Li<sup>2</sup>, Gamaleldin Elsayed<sup>3</sup>, and Mohamed Elhoseiny<sup>1</sup>

<sup>1</sup> King Abdullah University of Science and Technology

<sup>2</sup> HealthUnity

<sup>3</sup> Google DeepMind

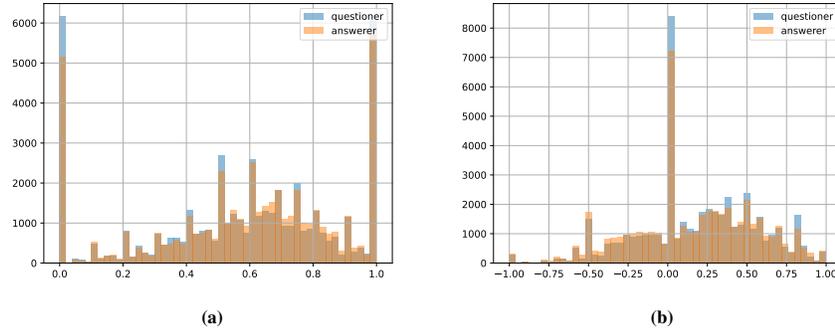
## 1 Contents

- Project Page <https://affective-visual-dialog.github.io/>
- Additional Details on Data Collection 2
- Data Inclusion / Exclusion Criteria 3
- Additional Dataset Analysis 4
- Implementation Details 5
- Details about LLMs and Vision-LLMs 6
- Emotion Guidance with Answers 7
- Human Studies Details 8
- More Examples from Affective Visual Dialog 8
- Additional Quantitative Results for Dialog-based Q&A task 9.1 and Explanation Generation task 9.2
- Additional Qualitative Results for Dialog-based Q&A task and Affective Explanation Generation task 10
- Example dialogs based on real images 11

## 2 Additional Details on Data Collection

**Visual Stimuli.** Our study employed WikiArt’s [6] artwork as a visual aid to elicit emotional responses. We carefully selected artworks that received both positive and negative evaluations from ArtEmis v1 [4] and v2 [9]. More than 60,000 artworks met our inclusion criteria for having both positive and negative emotional attributions and explanations. To maintain consistency in presenting the visual stimuli, we scaled down the largest image size to 600 pixels while preserving the original aspect ratio. This scaling procedure aimed to reduce the loading and scrolling time required for higher-resolution images, and to ensure a uniform presentation of visual stimuli. Furthermore, we sought to expand our visual stimuli by including real images in our study. We curated more than 580 images from [3] and applied the same data collection process as we did with the artworks. Examples of these images can be found in Figure 11.

**IRB protocol.** As our study involved human participants, we adhered to the IRB protocol (Protocol number 21IBEC049) and obtained informed consent from interested



**Fig. 1:** Distribution of different scores of explanations from Questioner and Answerer a) subjectivity scores (closer to 1 means the explanation is more subjective) b) sentiment scores (0 means neutral tone, -1 means the explanation conveys negative mood, and 1 vice versa)

individuals prior to their participation. Specifically, participants were presented with a consent form in the form of an onboarding task before beginning the study, which outlined the research purpose, participant requirements, potential risks, personal identity protection (if applicable), and compensation details.

**Interfaces.** Our data collection was conducted using the MTurk crowdsourcing platform. We built on top of the Mephisto framework [15] to create customized front-end interfaces for data collection. The interfaces utilized for both the questioner and answerer are illustrated in Figure 12, and the video-capturing chat process can be found inside provided zip file for supplementary.

**Instructions.** Both Questioner and Answerer were asked to follow the following rules (see Figure 13):

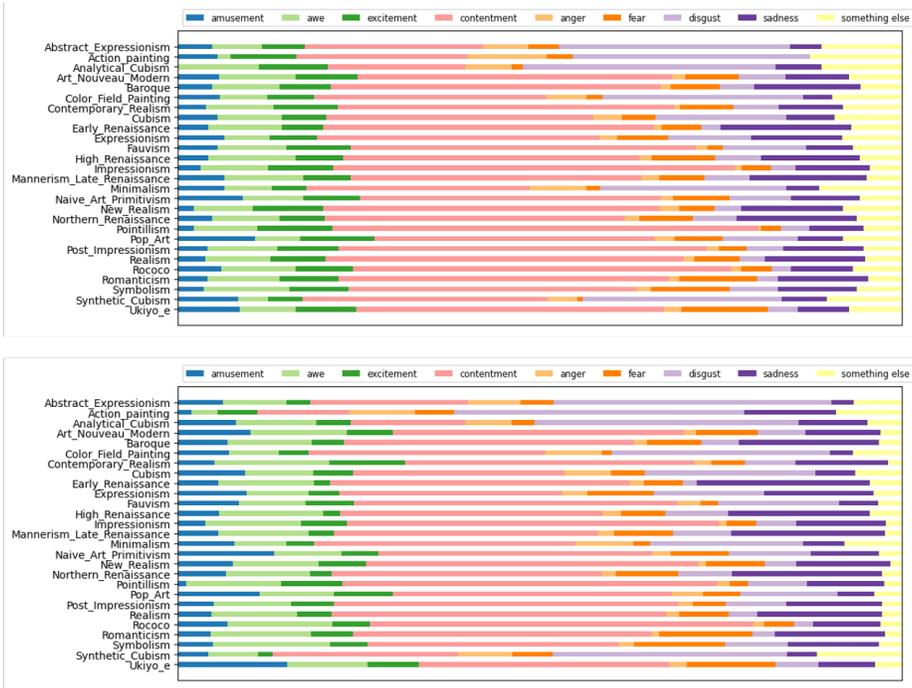
- to directly start the conversation and not make small talk
- not to write potentially offensive messages
- not to have conversations about something other than the image
- to just either ask questions or answer questions about an image (depending on assigned the role)
- not to use chat/IM language (e.g, "r8" instead of "right").
- to use professional and grammatically correct English.
- to have a natural conversation
- to not ask the answerer to provide his/her feelings about image explicitly

### 3 Data Inclusion / Exclusion Criteria

In order to ensure the inclusion of high-quality dialogs, we employed a strict set of inclusion criteria. Specifically, we only included dialogs that contained the full 10 turns, with both the Questioner and Answerer providing their emotional explanations regarding the hidden image. From the initial 107,912 dialogs collected, 17,435 were deemed



**Fig. 2:** Wordcloud of common words of **AffectVisDial**. The size of each word is proportionate to its frequency from a) explanations before observing the image; b) explanations after observing the image; c) explanations of the answerer.



**Fig. 3:** Emotion Distribution among different genres for: top) questioner; bottom) answerer

*“incomplete”* and excluded due to having less than 10 turns. Additionally, we excluded dialogs that contained inappropriate or irrelevant content that deviated from our given instructions, such as offensive messages or chitchat. After manual inspection, 40,477 complete dialogs were excluded for noncompliance with our guidelines, leaving a total of 90,477 dialogs. To promote productive and insightful dialogue that delved deeper into exploring the hidden image, Questioners were instructed to avoid asking redundant questions that could easily be answered by referring to the given opinions. Our careful selection and filtering of data was intended to ensure the quality and utility of the **AffectVisDial** dataset.

## 4 Additional Dataset Analysis

It is noteworthy that the language used in emotional explanations has an affective nature. To illustrate this point, we utilized a sentiment analyzer (TextBlob [2]). The distribution of subjectivity scores for emotional explanations provided by both Questioners and Answerers can be seen in Figure 1a, where a score closer to 1 indicates a higher level of subjectivity in the explanation. In addition, Figure 1b displays the distribution of sentiment scores, where a value of 0 signifies that the explanation conveys a neutral tone. From these figures, we can observe that explanations are subjective and sentimental in nature. Figure 2a, 2b, 2c shows the wordcloud of common words in **AffectVisDial**. Figure 3 shows the emotion distribution among different artistic styles for both Questioner and Answerer. It can be seen that the most dominant emotion across all genres is “*contentment*”.

## 5 Implementation Details

**Visdial-BERT** We follow the official implementation of Visdial-BERT [10] and adapt it to our dataset. Specifically, we modify the setting to classification by making the model select the most probable correct answer from a list of 100 candidate answers. The 100 candidate answers are randomly selected from original answers in our dataset. We experiment with 5 different sets of randomly selected answers and report the average of 5 runs in the visdial-bert table. The training setting and hyper-parameter choice are the same as in the original paper [10].

**LTMI [11]:** Following official implementation of LTMI<sup>4</sup>, we detect  $K = 100$  objects from each artworks in our dataset. We build a vocabulary of size 25,815 words that appear at least five times in the training split for the question and history features. The captions, questions, and answers are truncated or padded to 40, 20, and 20 words, respectively. We use pre-trained 300-dimensional GloVe vectors provided by authors to initialize the embedding layer. The embedding layer is shared for all the captions, questions, and answers. We train this model on our dataset using the Adam optimizer with 30 epochs. The learning rate is warmed up from  $1 \times 10^{-5}$  to  $1 \times 10^{-3}$  in the first epoch, then halved every 2 epochs. The batch size is set to 32.

**NLX-GPT [13]:** NLX-GPT is a language model that can simultaneously predict an answer and its corresponding explanation. We adapt it into our visual dialog setting to predict the emotion and corresponding emotion explanation. We make the question input as ‘*What is the emotion?*’ and format the emotion-explanation prediction as ‘*I feel **EMOTION** because **EXPLANATION***’. From questioner perspective, the *emotion\_before* label is the emotion given by questioner before getting access to image, thus we remove the visual backbone of NLX-GPT (i.e., set *add\_cross\_attention* as False) to follow the same logistic of collected dataset. We follow the official implementation<sup>5</sup> and benchmark it on our proposed dataset. The original maximum sequence length is only 70, to make it fit in our setting with long dialog input, we set *max\_seq\_len* to 400. We train this model for 100 epochs using AdamW optimizer with learning rate  $1e-5$  and

<sup>4</sup> <https://github.com/davidnvq/visdial>

<sup>5</sup> <https://github.com/fawazsammani/nlxgpt>

select the model with the best emotion prediction F1 score to evaluate on test set. All experiments are conducted on 4 NVIDIA V100 GPUs with batch size 32.

**BART-Large and T5-Large:** We experiment with both questioner and answerer explanation generation setup. In both setups, the maximum sentence length is 350 and the maximum generated sentence length is 50. BART-large is trained for 25 epochs with a batch size of 32 and a learning rate of  $1e-5$ . T5-Large is trained for 5 epochs with a batch size of 16 for 5 epochs on 4 NVIDIA A6000 GPUs.

## 6 Performance of LLMs and Vision-LLMs

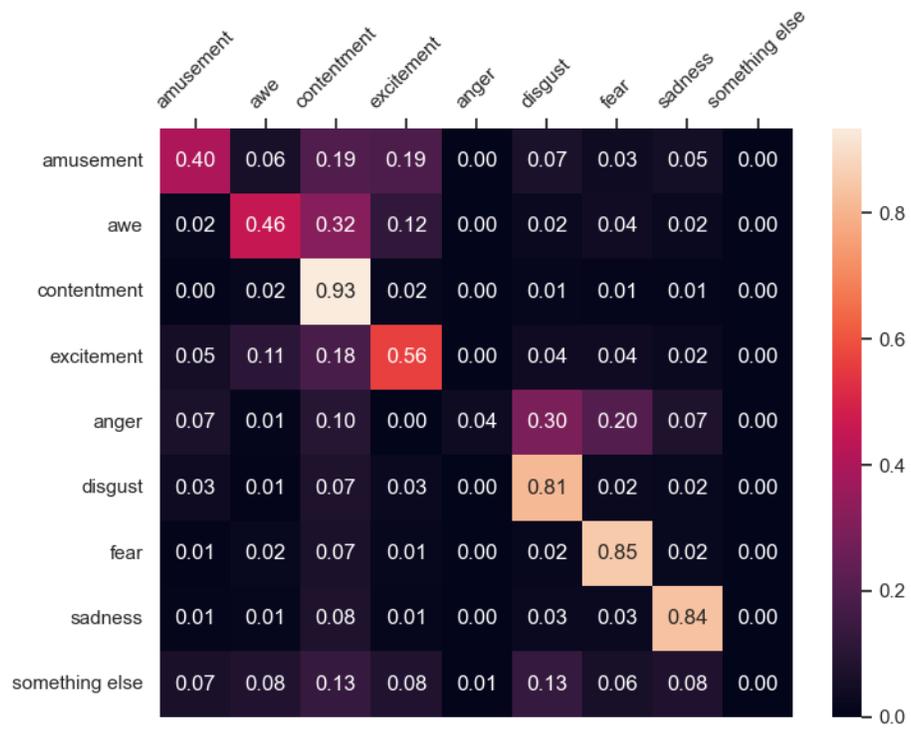
**Zero-shot Evaluation:** To evaluate the zero-shot capabilities of both language and multimodal foundational models, we prompted LLaMa2-7b-chat [14], GPT-4 [1], and MiniGPT-4-v2 [5] with the following instructions: "What do you feel after reading this text? Choose one of the following emotions: excitement, sadness, anger, contentment, something else, disgust, fear, amusement, and awe. Explain why you feel this way:  $[E_1]$   $[C_1]$  and  $[E_2]$   $[C_2]$  and  $[D]$ . Choose only one emotion, and do not repeat dialogue. Respond with 'I feel ... because ...'." Here,  $E_1$  and  $E_2$  represent opposing emotions,  $C_1$  and  $C_2$  are corresponding opposing opinions, and  $D$  represents a 10-turn dialogue.

## 7 Emotion Guidance with Answers

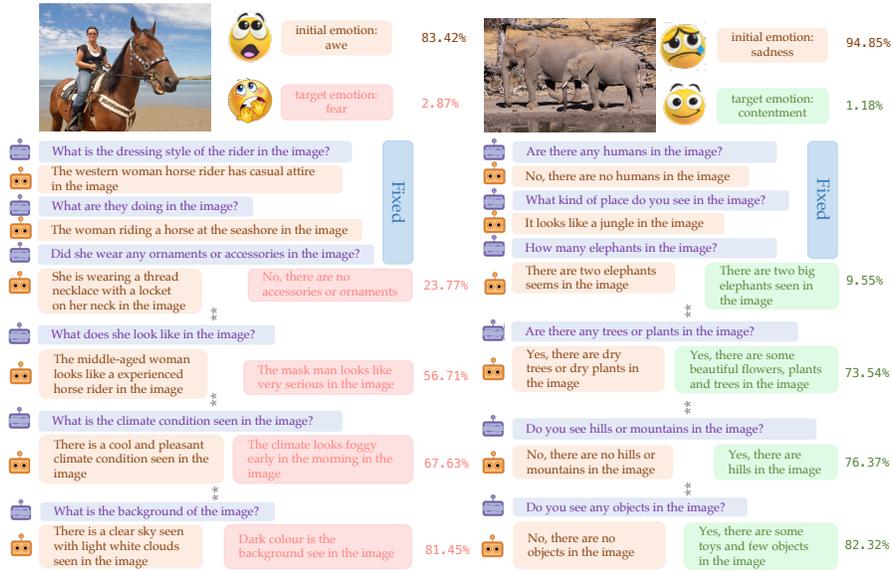
In this section, we present a methodology to associate emotions with answers in the context of dialogues.

**RoBERTa-based Emotion Classifier.** We begin by fine-tuning a pre-trained RoBERTa model [8] as an emotion classifier on our proposed dataset, following the approach described in [9]. The resulting confusion matrix for dialog-based emotion classification using this RoBERTa-based classifier is presented in Figure 4. Each column shows how percentage-wise the model confuses the specific emotion with all available emotion classes. Each row sums to 1. Each row of the matrix sums to 1. Notably, the highest confusion rate is observed among emotions of the same sentiment (positive, negative). It is worth noting that the least frequently occurring emotion class in **AffectVisDial**, i.e., anger, is also the most frequently misclassified one.

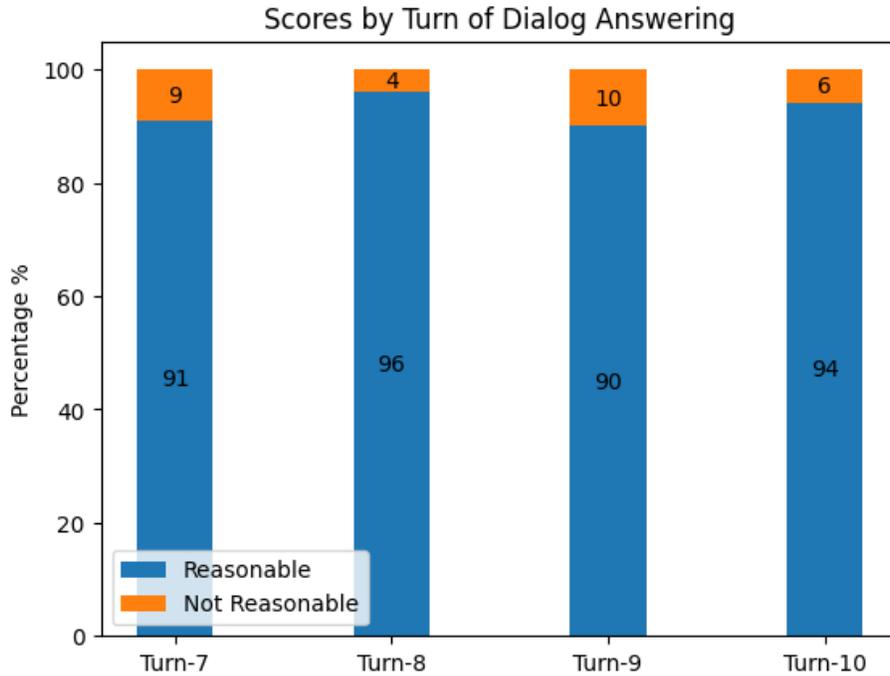
**Achieving the Target Emotion.** To guide the predicted emotion towards a targeted opposing emotion, we gradually change the answer turn by turn, starting from the third turn of the dialogue. We select candidate answers that are similar to the current question and choose the answer that yields the highest prediction probability for the target emotion. This process continues until the target emotion is achieved. Here we assume that with the guidance of emotion classifier, we can achieve the desired emotion. We visualize the effectiveness of this approach through two examples for real images in Figure 5, which demonstrate how the prediction probability for the target emotion increases as we gradually replace emotion-related answers.



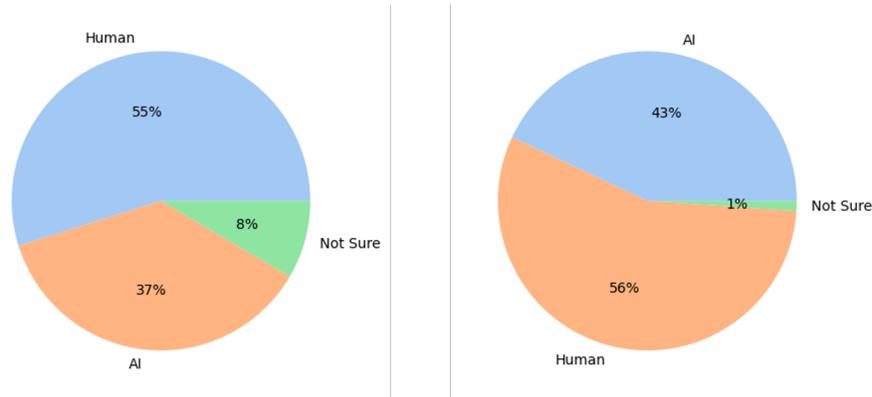
**Fig. 4:** Confusion matrix for dialog-based classification of emotion.



**Fig. 5:** Guiding emotion by altering answers. We show the original and replacing answer and corresponding prediction probability of target emotion by turns.



**Fig. 6:** Results of Reasonableness Test.



**Fig. 7:** Results of Turing Test. More than 50% of generated explanations are considered human-like

## 8 Human Studies details

In order to validate the performance of our model, we conducted human studies on Amazon Mechanical Turk.

**Reasonableness Test:** This study sought to investigate the effectiveness of a model in answering questions based on dialog and the provided image. Specifically, we selected a subset of 100 dialogs from our dataset and asked five participants to evaluate the reasonableness of the model’s answers to follow-up questions in each dialog. This study was conducted four times, focusing on the last four turns of each dialog. The results indicated that over 90% of the model’s answers were deemed reasonable based on the majority vote of participants (as depicted in Figure 6). The human study interface utilized for this evaluation is illustrated in Figure 14.

**Turing Test:** In this study, we asked participants to indicate whether the produced emotion and corresponding explanations from explanation generation models are human-like or not. Specifically, we posed the question “Was the emotion explanation generated by a human or the AI system”. We did this study with two variants: a) explanations based on solely dialog and b) explanations based on an image and corresponding dialog. The user interfaces for both variants are provided in Figure 15. Results show that more than 55 % of explanations are considered human-like (see Figure 7).

## 9 Additional Quantitative Results

### 9.1 Dialog-based Answering

Table 1 shows the performance of LTMI-D [11] model after a certain number of turns. As the dialog progresses, the model performance increases in all metrics, indicating the importance of the dialog history to achieve good results.

| Turn | R@1(↑) | R@5(↑) | R@10(↑) | MRR(↓) | MR(↓) |
|------|--------|--------|---------|--------|-------|
| 0    | 0.005  | 0.009  | 0.015   | 0.99   | 45.67 |
| 2    | 0.01   | 0.019  | 0.025   | 0.87   | 35.48 |
| 4    | 0.1    | 0.15   | 0.3     | 0.71   | 27.58 |
| 6    | 0.16   | 0.21   | 0.57    | 0.65   | 19.47 |
| 8    | 0.2    | 0.27   | 0.69    | 0.57   | 10.98 |
| 10   | 0.23   | 0.35   | 0.79    | 0.47   | 6.1   |

**Table 1:** Performance of LTMI-D [11] after every 2 turns on Affective Visual Dialog task.

| Model          | Image | Emotions | Captions | Dialog | BLEU(↑) | BERT(↑) | BART(↓) | Emo-F1(↑) |
|----------------|-------|----------|----------|--------|---------|---------|---------|-----------|
| NLX-GPT [13]   | ✓     | ✓        | ✓        | ×      | 0.08    | 0.63    | -6.97   | 26.62     |
| NLX-GPT [13]   | ✓     | ✓        | ✓        | ✓      | 0.16    | 0.75    | -6.59   | 46.27     |
| BART-Large [7] | ✓     | ✓        | ✓        | ×      | 0.003   | 0.26    | -5.57   | 17.47     |
| BART-Large [7] | ✓     | ×        | ✓        | ✓      | 0.19    | 0.66    | -4.51   | 41.44     |
| BART-Large [7] | ✓     | ✓        | ✓        | ✓      | 0.19    | 0.65    | -4.51   | 42.49     |
| T5-Large [12]  | ✓     | ✓        | ✓        | ×      | 0.015   | 0.34    | -5.50   | 22.13     |
| T5-Large [12]  | ✓     | ×        | ✓        | ✓      | 0.17    | 0.64    | -4.69   | 36.31     |
| T5-Large [12]  | ✓     | ✓        | ✓        | ✓      | 0.18    | 0.65    | -4.66   | 38.05     |

**Table 2:** Results on emotion and explanation generation setup for Answerer.

## 9.2 Explanation Generation/Emotional Reasoning for Answerer

The outcomes of the experiment on the answerer’s emotion and explanation generation are presented in Table 2. The results indicate that incorporating dialog  $D$  as a part of the input significantly improves model performance. Specifically, BART-Large with  $D$  achieved better performance than BART-Large without  $D$  with an increase of 0.19 in BLEU, 0.39 in BERT, 1.06 in BART scores, and 25% in emotion F1. The same trend was observed for T5-Large, where models trained with  $D$  outperformed their counterparts by 0.17 in BLEU, 0.31 in BERT, 0.84 in BART, and 16% in emotion F1.

## 10 More Qualitative Examples

More examples from our dataset can be seen in Figure 8. Figure 9 shows outputs from baselines in Dialog-Based Q&A task at different turns. More generated explanations can be seen in Figure 10. Some examples collected on real images are shown in Figure 11.



The blended colors of the child's skin is very unique.

The little girl looks lonely and the muted colors make me think of gloomy weather.

How many people can be seen in the image?  
I can see only one young girl in the image.

What is the girl doing?  
The girl is just standing in the middle of the market.

How old does the girl look like?  
The girl looks like 14 to 15 years old.

What clothes is the girl wearing?  
She is wearing casual winter outfit.

What is the weather in the image?  
The weather is like close to the winter.

What time of the day is shown in the image?  
It looks like sunset.

Does the market look like it is situated in the town or village area?  
The market looks like it is situated in the village.

What is the color scheme used in the image?  
Its tragic color scheme.

What expression can be seen on the face of the girl?  
Its neutral.

What type of shops can be seen in the market?  
Cycle repair and fruit shops.

**Emotion explanations:**  
I feel sad for the young girl as she is alone in the market and wonder if she could be lost.

I feel awe because of the pretty features used to paint the girl and intrigued as to what she is thinking.

A young girl is standing in the market and she looks very beautiful and charming, and I think she is selling oranges in the market at a young age which looks like she has so many responsibilities so she is selling fruits so it gives me a proud and good feeling while seeing this responsible girl.



The meadow is foggy and pleasant in the early morning light.

The sky looks like it's really muddy since it seems to be glowing with an odd shade of gray.

Is this an indoor or an outdoor scene?  
It is an outdoor scene.

Is it rural or is it urban?  
It looks to be rural.

Are there any people present?  
No, there are no people.

Are there any animals?  
No, no animals.

What is the weather like?  
It is a little hard to tell because of how the image is painted, but it looks to be either sunny or cloudy.

Is the sun itself actually visible?  
No, the sun is not visible.

Is there anything that can date the picture?  
Not that I can see.

Can you tell what time of day it is?  
Sometime during the day when the light is shining, May be midday.

What are the main colours here?  
Purple, green, and blue.

Does it put you in mind of any particular painter?  
Van Gogh, maybe.

**Emotion explanations:**  
This sounds as if it is entirely non threatening, just a peaceful outdoor scene.

Some of the shapes are frankly rather creepy ... like hands stretching out of the earth.

This painting makes me think of having a picnic in the forest.



This looks like Merlin the wizard! I love reading about King Arthur when I was younger.

He appears to be casting an evil spell.

How many people do you see in the image?  
There is only one person seen in the image.

Do you see any spiritual beings in the image?  
No, there are no spiritual beings in the image.

Do you see any houses or buildings in the image?  
No, there are no houses or buildings seen in the image.

Do you see plants or trees in the image?  
Yes, there are some plants seen in the image.

What is the person in the image doing?  
The person seems like he is sitting on a ledge in the image.

What type of dress does this person wear?  
He seems to be wearing a thin long robe.

Does he have long hair with a long nose in the image?  
Yes, he has long hair with a long nose in the image.

Does the person in the image seem young or old?  
He seems to be an adult and he is not too young and not too old.

Does the man have a staff or a stick with him in the image?  
He does not have any staff or a stick with him.

What does the background look like?  
The background is black with the name printed on it and it is merlin.

**Emotion explanations:**  
The man sitting on a ledge with a thin robe and it seems to be night as well as his hair and nose make me wonder if the person is a witch and it gives me creeps and hair down my spine.

The man looks kind of creepy but he looks normal and is taking a rest there seems nothing wrong with the person and I feel nice as he seems like a hard worker and a poor guy working hard and taking rest over a ledge.

The man is sitting at the edge while wearing a long robe, and she looks young but his facial expressions are evil because his eyes and his nose remind me of a witch who has a lot of magical powers so it makes me feel scared of him.



The river is bright and pretty.

We always take good thing to eat with the bad thing.

Do you see humans in the image?  
No, there are no humans in the image.

What kind of weather is presented in the image?  
The weather looks calm and pleasant in the image.

What is the background of the image?  
The sky is in the background of the image.

Do you see any boats in the river?  
It seems like there is a boat in the river in the image.

How many colours are there in the image?  
There are a lot of colours in the image.

What is the most dominant colour in the image?  
White is the most dominant color in the image.

Do you see animals in the image?  
No, there are no animals in the image.

Do you see any houses or buildings in the image?  
Yes, there is a building in the image.

Is there any signs or signatures in the image?  
No, there are no signs or signatures in the image.

Are there hills or mountains in the image?  
Yes, there is a mountain in the image.

**Emotion explanations:**  
The river there is very calm and the atmosphere is also quiet, which scares me because there are no people in the image.

What I had imagined went poor as soon as I saw the image. This kind of feeling makes me feel nauseous.

The builders built an awesome building beside the river in good and calm weather with amazing scenery, the engineers are planning really nicely because they want this place to become a tourist spot, and they want to attract the tourist with its amazing beauty, it is really an awesome thing, it makes me feel amazing.

Fig. 8: More examples from AffectVisDial.

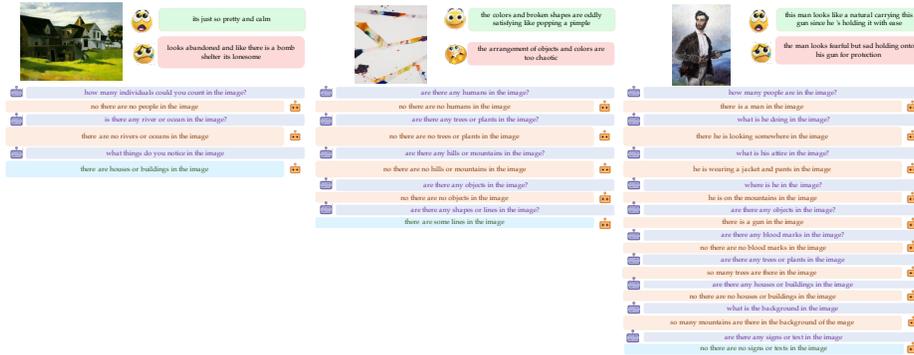


Fig. 9: Outputs from baselines in Dialog-Based Q&A task at different turns. Given the image, two opinions, and dialog context, the model outputs an answer (highlighted in green text).



Fig. 10: Predicted Emotions and Generated Explanations from the baselines.



that assortment of seasoning looks really delicious i would want to try a bite of that 😋

i am not in favour of unhealthy foods 😞

**Emotion explanations:**

The man is eating delicious street food to satisfy his hunger, it is really a bad habit because he is eating unhealthy food and he will get health issues from the street food because the street food hasn't been cooked in hygienic that's why he gets sick from this street food, it makes me feel sad about street food. 😞

The food item looks very delicious and there is some onion garnish along with some tomato ketchup on it and it looks tastier rather than normal and I like to eat such fried item because they're very tasty but it is not good for our health so I eat them once in a month which gives me a awesome feeling. 😋



i do like the color of the truck i am wondering if it is for business or possibly a taxi 😊

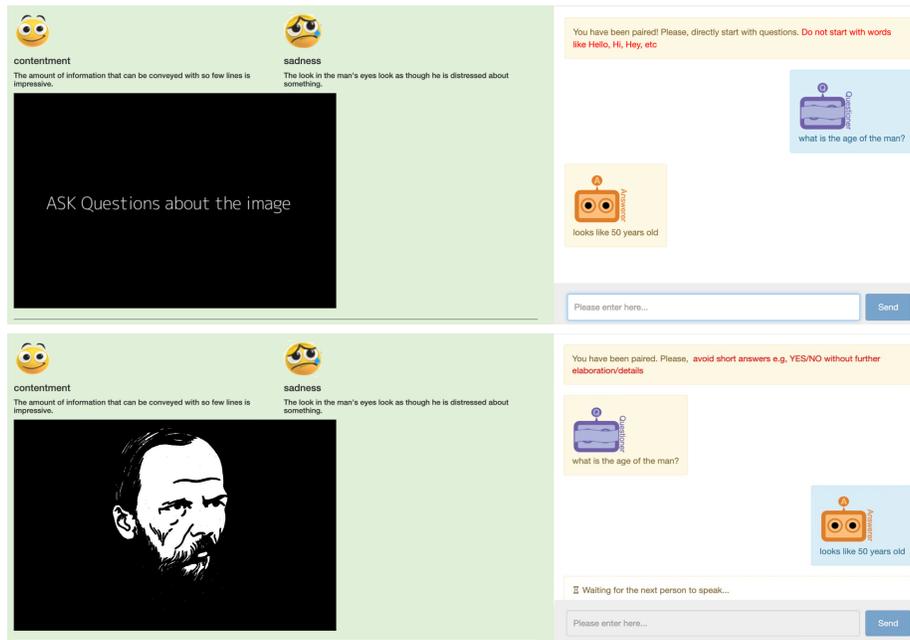
because it feels like a police car and even thinking about the police driving around my neighborhood gives me great sadness for my community 😞

**Emotion explanations:**

The yellow color police car is attractive and the government is protecting people's lives by putting signals and rules, the police are protecting the people. It makes me feel good because the government protects people's lives with their rules. 😊

The yellow colour police car standing beside the road because the cops are controlling the traffic and the cops make people follow the traffic rule, it is really a nice thing because they are thinking of civilian's safety on the road, it makes me feel good because they make a lot of rule to drive the vehicle on the road very safely. 😊

Fig. 11: Example dialogs based on real images.

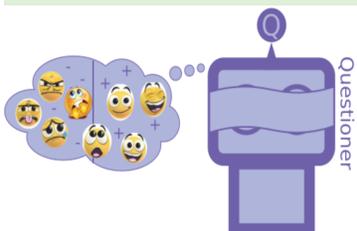


**Fig. 12:** During our data collection, we employed a specific interface for both the Questioner and Answerer. The user interface utilized by the Questioner is displayed in the top image, while the Answerer's interface is shown in the bottom image. It is important to note that only the image was visible to the Answerer during the data collection process.

Both the Questioner and Answerer should follow the following instructions during the conversation

- Please directly start the conversation. Do not make small talk.
- Please do not write potentially offensive messages.
- Please do not have conversations about something other than the image. Just either ask questions, or answer questions about an image (depending on your role).
- Please do not use chat/IM language (e.g., "r8" instead of "right"). Please use professional and grammatically correct English.
- Please have a natural conversation. Unnatural sounding conversation including awkward messages will be rejected.
- Please note that you are expected to complete and submit the hit in one go (once you have been connected with a partner). You cannot resume hits.
- You have maximum of 3 minutes response time per turn before your HIT is rejected
- Questioner should keep his or her questions about the content of the image and do not ask the answerer to provide his or her feelings about the image explicitly.

Please complete one HIT before proceeding to the other. Please don't open multiple tabs, you cannot chat with yourself. We may measure the level of engagement in the task and the task may terminate if a low level of engagement is detected or if instructions are violated. Your data will be recorded ONLY when the dialog is complete (i.e., reaching a decision after 10 Questions and 10 answers) and no violation of instructions is detected. Thus, we expect the questioner and answerer to work as a team to fully complete the HIT and receive the payment. Your participation is voluntary and you can stop at any time, but you will be paid for completed tasks only.



**Role of Questioner**

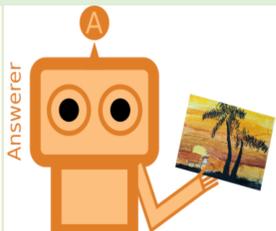
The questioner asks questions about a piece of art that is visible only to the Answerer. The Questioner's task is to decide on an emotion/feeling based on the responses.

**Instructions to the Questioner**

In this task, to help you understand the task, imagine that you are a blind person who wants to engage in an emotional experience about an artwork and form an opinion about it by querying another person. As a "Questioner", you do not have access to the art image, but you will have access to two subjective descriptions reflecting different emotions about the image. You will engage in a dialog with a person (a fellow Turker) who has access to the artwork image. Your role is to ask questions that are specific to the content of the image to decide on an emotion that the hidden artwork may trigger for you. At the end of the conversation, you will be given 9 possible emotions (awe, contentment, excitement, amusement, sadness, anger, disgust, fear, or something else) to choose from based on your conversation with your fellow Turker. At the end, you will also be requested to provide a text description that explains based on the conversation why you chose the selected emotion. Please refer to pieces of information from the conversation that informed your decision. **Very generic questions** (e.g., "What is the image about?", "you are welcome", "no more questions?") are not allowed.

Common bad examples for the **Questioner** that leads to HIT rejection:

- Very generic questions: What is the image about? What is depicted?
- Ask about the feelings of the answerer: How do you feel about the image?
- Irrelevant questions/chitchat: I am good, weather is great. How are you doing?
- Offensive language: ...



**Role of Answerer**

The Answerer provides answers to the questions about the content of a visual artwork.

**Instructions to the Answerer**

In this task, to help you understand the task, imagine that you are helping a blind person explore and appreciate an artwork that you only can see by providing answers about the content of the artwork. As an "Answerer" you will have access to the artwork image as well as two subjective descriptions reflecting different emotions about the image. You will engage in a dialog with a person (a fellow Turker) who can not see the artwork image. Imagine if you are helping a blind person to experience the artwork that you can see but he/she does not. Your role is to help answer questions about the visual content that is specific to the art piece (not general questions). Also, please help the presumably blind fellow turker to form his/her own opinion about the art piece without imposing certain emotions, e.g. just saying this image is sad or joyful in your answers. Please also provide detailed answers (e.g., do not use short answers such as yes/no/maybe). What is important here is to help him/her create an emotional experience about the artwork. Please focus on describing the content from an artistic point of view including textures, the colors, peoples, animals, etc. Please keep the following in mind while chatting with your fellow Turker:

- Please keep the following in mind while chatting with your fellow Turker:
- Short answers: yes, no, maybe, .....
- Provide emotions: picture is depressing as an answerer.
- Irrelevant answers/chitchat: I am good, weather is great. How are you doing?
- Offensive Language: ...
- Please have a natural conversation. Unnatural sounding conversation including awkward messages will be rejected.

Fig. 13: Instructions for both Questioner and Answerer

Please, decide whether the answer is correct.

### Instructions - is the answer correct?

1. You will be given a part of a conversation of two people about the image. The person who asks question cannot see image but rather two opinions about the image. The other person who provides answer can see the image and also two opinions.
2. Your task is to decide whether the person gave correct answer highlighted in blue color given an image.



Two opinions:

the shadows are swallowing this painting as it gets darker and the stormy clouds roll in making this a dramatic painting

the clouds coming up behind the hills and the church is ominous

Dialog:

do you see any people in the image

no there are no people in the image

are there any objects in the image

yes there are objects in the image

what are the objects in the image

the trees along with the lake and houses in the image

what is the background of the image

the background seems dark clouds in the sky in the image

what does look like the weather in the image

the weather seems to be cool and pleasant in the image

do you see any sign or text in the image

no there is no sign or text in the image

are there any trees or plants in the image

yes there are trees and plants in the image

do you see any hills or mountains in the image

no there are no hills or mountains in the image

Do you think the answer in blue color is reasonable to the given question based on dialog and image?

Reasonable

Not Reasonable

**Fig. 14:** User Interface for Reasonableness Test.

Please, decide whether the answer is correct.

**Instructions -**

- You will be evaluating whether the emotional explanations (highlighted in blue color) based on two opinions and the information from a conversation about some hidden image.
- Your task is to determine whether the emotion explanations comes from **humans** or **generated by the AI system**
- Please, read two opinions and dialog about the image carefully and decided whether the emotion explanations are from humans or AI

**Two opinions:**  
sadness: the person in the image seems like they are sad or upset which makes me sad  
awe: I am in awe as this painting has a beautiful woman reading a book, drinking tea and eating grapes. Reminds me of learning about old times.

**Dialog:**  
what are the most important details in this picture??  
In the picture, I see a woman reading a book and some vegetation  
would you describe me the place in the picture??  
She is the bigger things I see and she is in the middle of the photo  
so would you tell what she is wearing??  
She is reading a book and eating tea  
is this woman inside or outside??  
She is outside, may be in a garden or in a peaceful place like a park  
what kind of clothes she is wearing??  
In general, she is wearing a white dress  
what is the facial expression of this woman??  
The woman looks very happy and seems to learn something interesting in the book  
what is the weather like in this image??  
All feel, the weather is very clear  
what else can you see around the woman??  
The woman is sitting in a chair in front of a table and on the table I see grapes  
how about the woman's hair??  
I can't see her hair because she is wearing a white hat  
so how is the environment in this image??  
I think that it is in a calm place, I see some trees and the environment is green

**Emotion Explanation:**  
contentment because I am in awe as this painting has a beautiful woman reading a book and eating grapes

Was the emotion explanation generated by a human or the AI system??

Human  
 AI system  
 Not Sure

Please, decide whether the answer is correct.

**Instructions -**

- You will be evaluating whether the emotional explanations (highlighted in blue color) based on two opinions and the information from a conversation about some hidden image.
- Your task is to determine whether the emotion explanations comes from **humans** or **generated by the AI system**
- Please, read two opinions and dialog about the image carefully and decided whether the emotion explanations are from humans or AI



**Two opinions:**  
sadness: the person in the image seems like they are sad or upset which makes me sad  
awe: I am in awe as this painting has a beautiful woman reading a book, drinking tea and eating grapes. Reminds me of learning about old times.

**Dialog:**  
what are the most important details in this picture??  
In the picture, I see a woman reading a book and some vegetation  
would you describe me the place in the picture??  
She is the bigger things I see and she is in the middle of the photo  
so would you tell what she is wearing??  
She is reading a book and eating tea  
is this woman inside or outside??  
She is outside, may be in a garden or in a peaceful place like a park  
what kind of clothes she is wearing??  
In general, she is wearing a white dress  
what is the facial expression of this woman??  
The woman looks very happy and seems to learn something interesting in the book  
what is the weather like in this image??  
All feel, the weather is very clear  
what else can you see around the woman??  
The woman is sitting in a chair in front of a table and on the table I see grapes  
how about the woman's hair??  
I can't see her hair because she is wearing a white hat  
so how is the environment in this image??  
I think that it is in a calm place, I see some trees and the environment is green

**Emotion Explanation:**  
contentment because the contentment at the woman is reading a book and eating grapes

Was the emotion explanation generated by a human or the AI system??

Human  
 AI system  
 Not Sure

**Fig. 15:** User Interface for Turing Tests for evaluating emotion explanations.

## References

- Openai api: Gpt-4. <https://platform.openai.com/docs/models/gpt-4> (2023), accessed: November 15, 2023
- Textblob: Simplified text processing. <https://github.com/sloria/textblob> (2023), accessed: March 7, 2023
- Achlioptas, P., Ovsjanikov, M., Guibas, L., Tulyakov, S.: Affection: Learning affective explanations for real-world visual data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6641–6651 (June 2023)
- Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11569–11579 (2021)
- Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning (2023), <https://arxiv.org/abs/2310.09478>
- Community: Wiki art. <https://www.wikiart.org/> (2020), accessed: 2020-11-06
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
- Mohamed, Y., Khan, F.F., Haydarov, K., Elhoseiny, M.: It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21263–21272 (2022)

10. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In: European Conference on Computer Vision. pp. 336–352. Springer (2020)
11. Nguyen, V.Q., Suganuma, M., Okatani, T.: Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In: European Conference on Computer Vision. pp. 223–240. Springer (2020)
12. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**(140), 1–67 (2020)
13. Sammani, F., Mukherjee, T., Deligiannis, N.: Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8322–8332 (2022)
14. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023)
15. Urbanek, J., Ringshia, P.: Mephisto: A framework for portable, reproducible, and iterative crowdsourcing (2023). <https://doi.org/10.48550/ARXIV.2301.05154>, <https://arxiv.org/abs/2301.05154>