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

Kilichbek Haydarov[1], Xiaoqian Shen[1], Avinash Madasu[1], Mahmoud Salem[1], Li-Jia Li[2], Gamaleldin Elsayed[3], and Mohamed Elhoseiny[1]

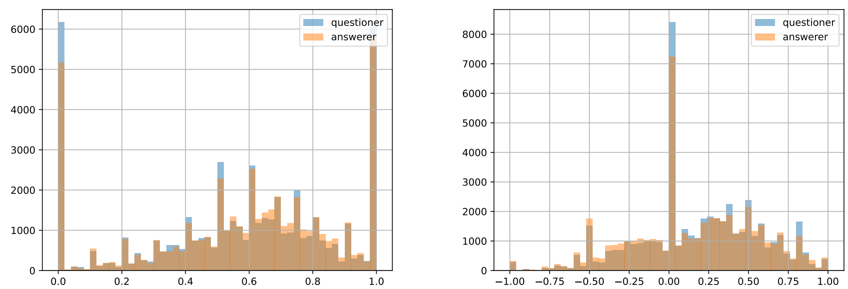
[1] King Abdullah University of Science and Technology
[2] HealthUnity
[3] Google DeepMind

1 Contents

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



(a) **(b)**

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

(a) (b) (c)

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.

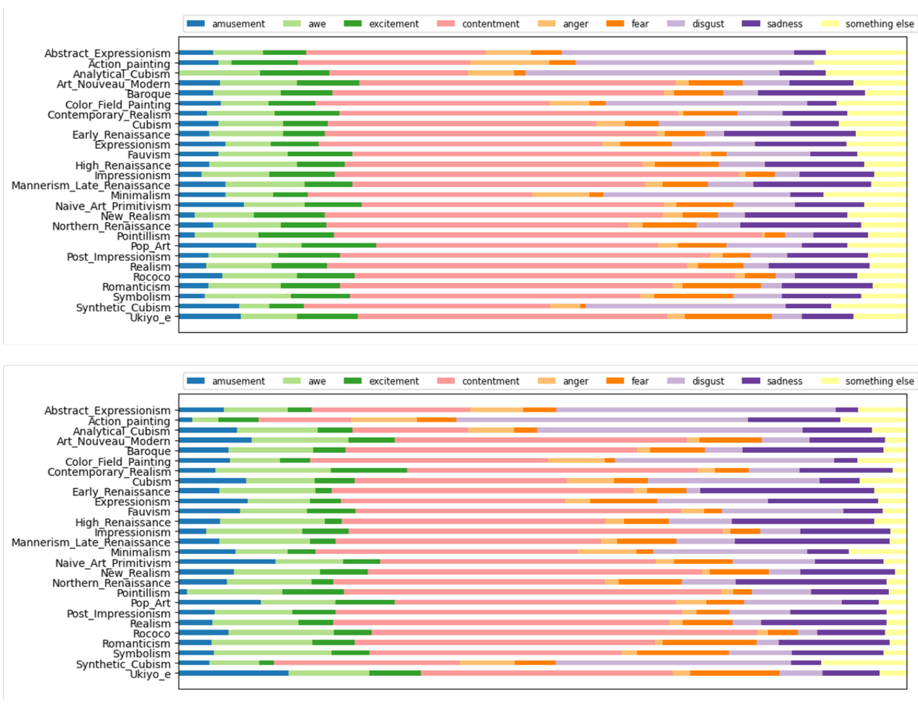


Fig. 3: Emotion Distribution among different genres for: top) questioner; bottom) answerer

"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 visualdial-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 [4], 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 1×10^{-5} to 1×10^{-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 *emtion_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[5] 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

[4] https://github.com/davidnvq/visdial
[5] 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.

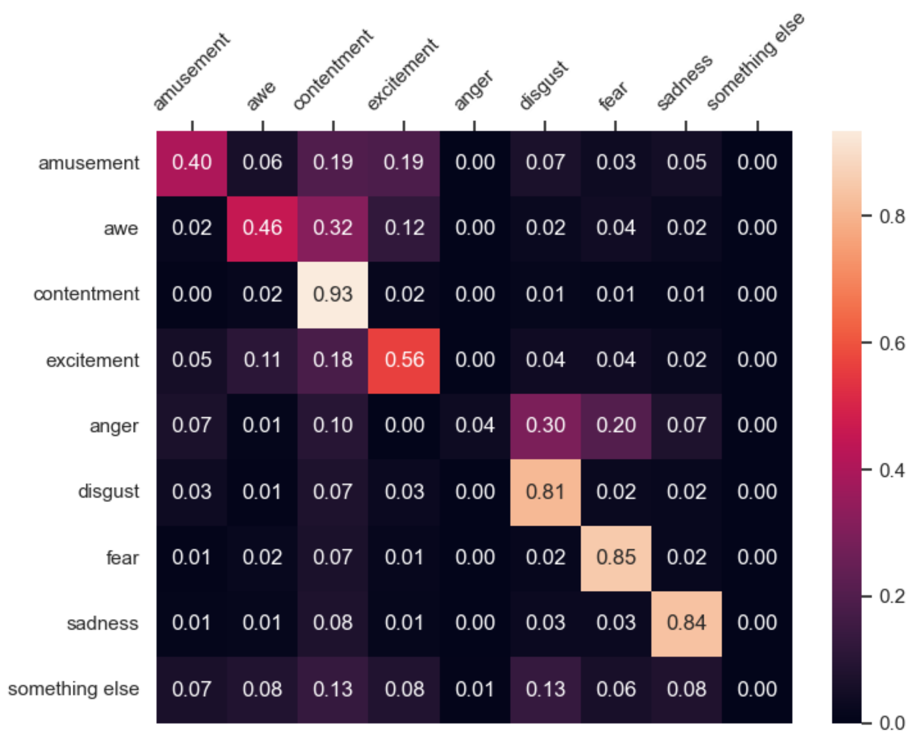


Fig. 4: Confusion matrix for dialog-based classification of emotion.

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

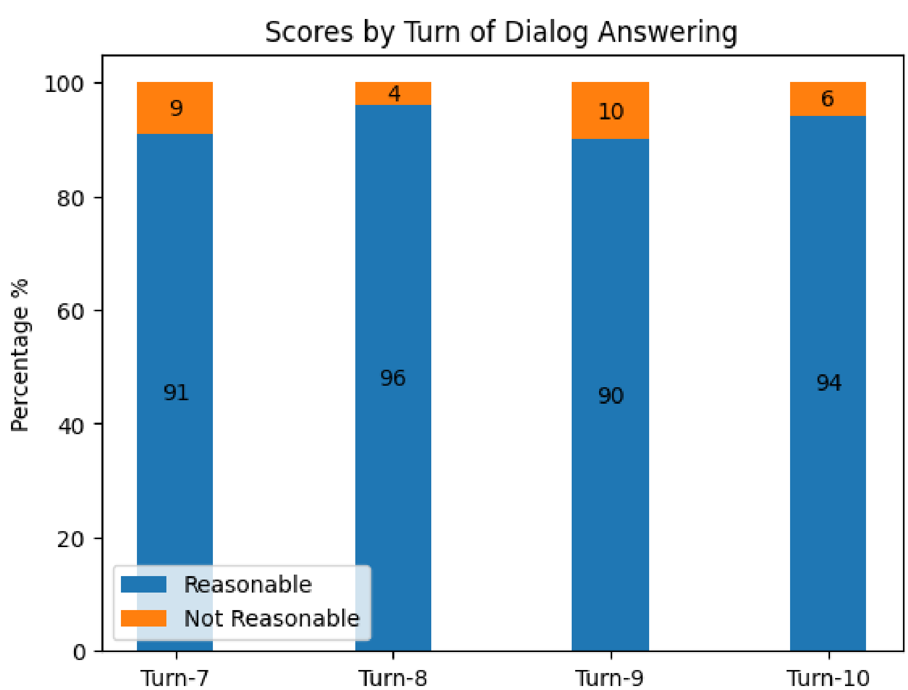


Fig. 6: Results of Reasonableness Test.

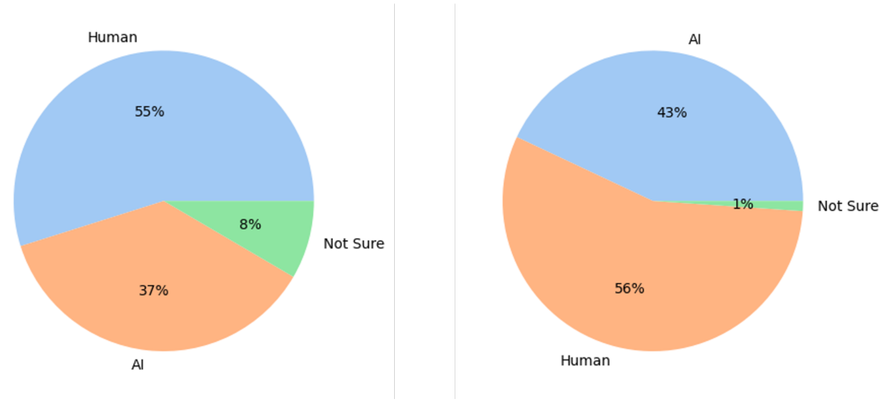


Fig. 7: Results of Turing Test. More than 50% of generated explanations are considered human-like

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.

Fig. 8: More examples from **AffectVisDial**.

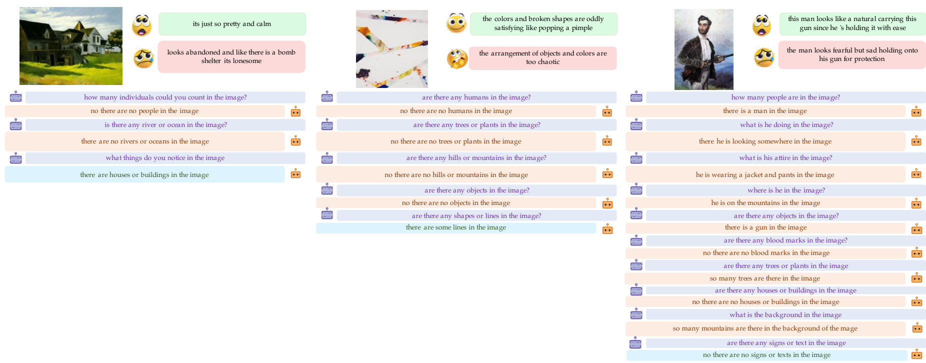


Fig. 9: Outputs from baselines in Dialog-Based Q&A task at different turns. Given the image, two opinions, and dialog context, the model outputs an answer (highlighted in green text).

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

Fig. 11: Example dialogs based on real images.

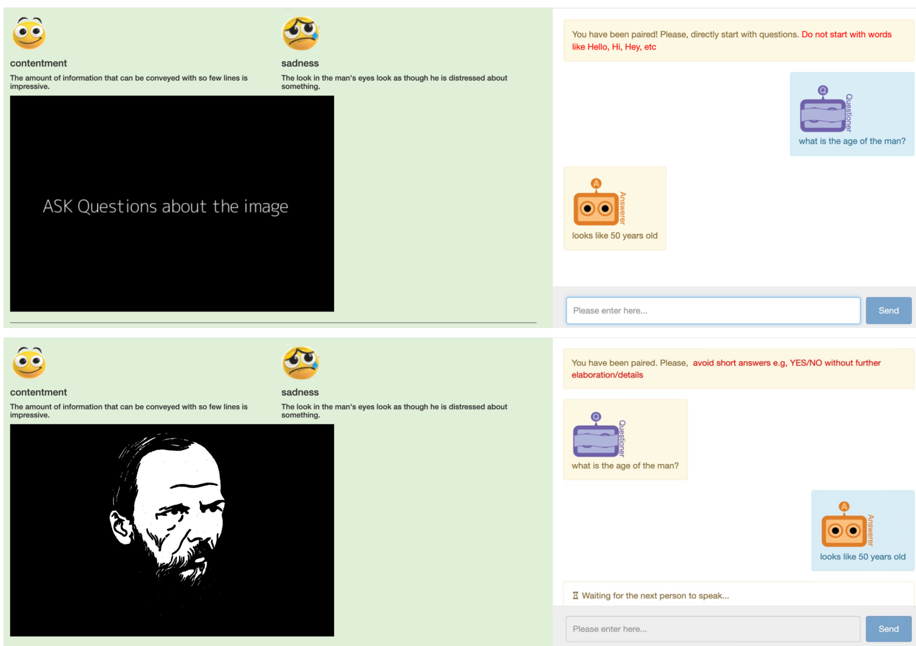
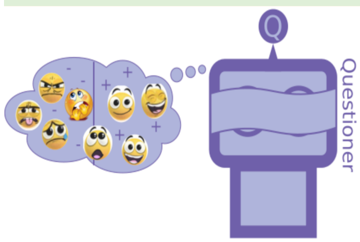


Fig. 12: During our data collection, we employed a specific interface for both the Questioner and Answerer. The user interface utilized by the Questioner is displayed in the top image, while the Answerer's interface is shown in the bottom image. It is important to note that only the image was visible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

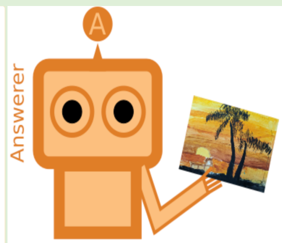
The questioner asks questions about a piece of art that is visible only to the Answerer. The Questioner's task is to decide on an emotion/feeling based on the responses.

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

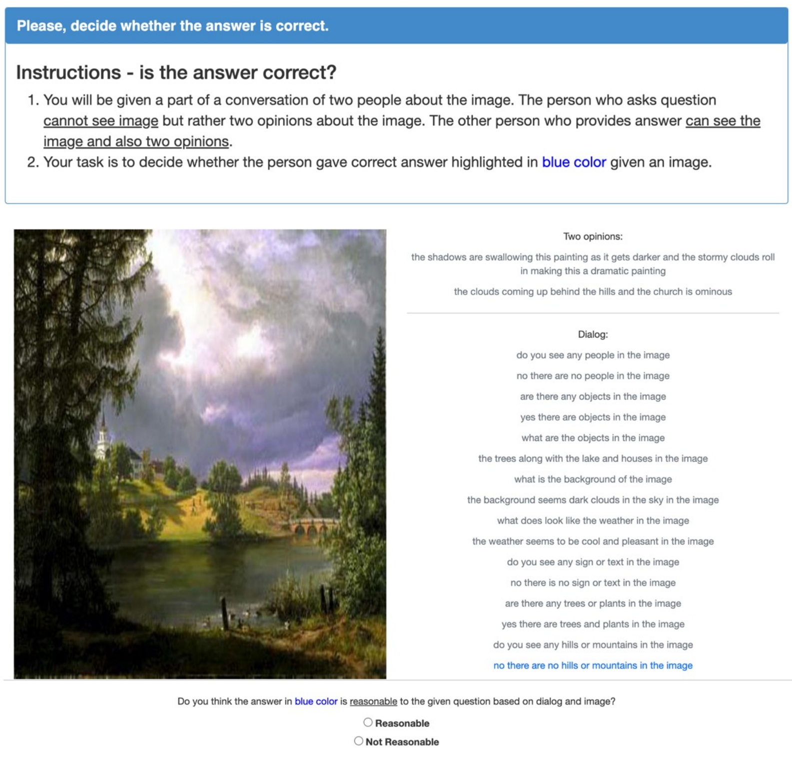


Fig. 14: User Interface for Reasonableness Test.

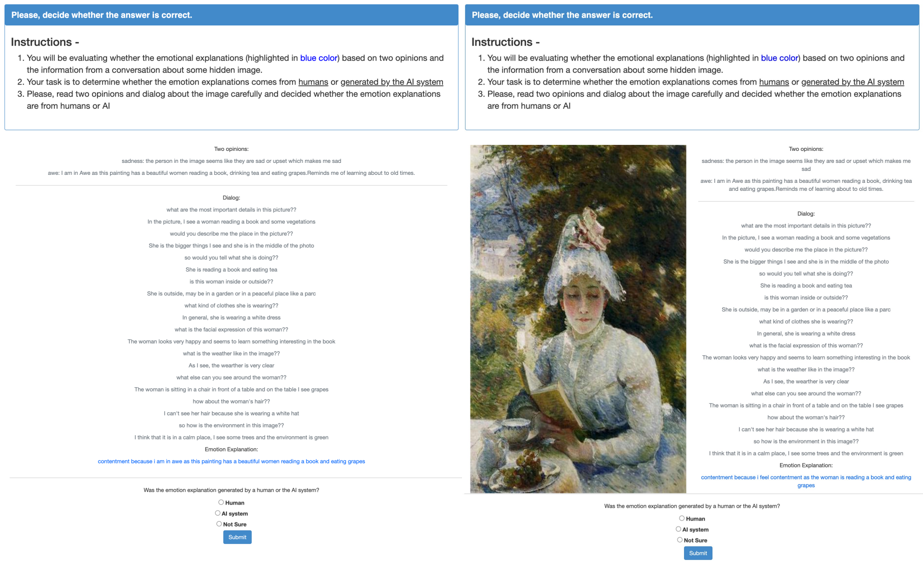


Fig. 15: User Interface for Turing Tests for evaluating emotion explanations.

References

1. Openai api: Gpt-4. https://platform.openai.com/docs/models/gpt-4 (2023), accessed: November 15, 2023
2. Textblob: Simplified text processing. https://github.com/sloria/textblob (2023), accessed: March 7, 2023
3. 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)
4. 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)
5. 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
6. Community: Wiki art. https://www.wikiart.org/ (2020), accessed: 2020-11-06
7. 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)
8. 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)
9. 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