It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection - Supplementary Material

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, Mohamed Elhoseiny
King Abdullah University of Science and Technology (KAUST)
{youssef.mohamed,faizan.khan,kilichbek.haydarov,mohamed.elhoseiny}@kaust.edu.sa

Supplementary Content:

This supplementary document include the following

Section 1: We elaborate on the results we reported in the main paper in Fig. 5. We briefly describe the semantic space theory and comment on why our combined dataset has better properties.

Section 2: We report the full set of results we obtained from training variations of SAT model on different training sets. We show how $Combined$ dataset results in superior models.

Section 3: We report results of extended linguistic analysis of our new dataset named $Combined$. We compare the values to the original ArtEmis dataset in order to highlight the advantages of contrastive data collection.

Section 4: We show different samples generated from neural speakers trained on the $Combined$ dataset.

Section 5: We include a link to the newly collected dataset as well as instructions and other statistics from Amazon Mechanical Turk about the data collection process. We also include a link to our implementation as well as more visualizations

1 Semantic Space Theory

Cowen and Keltner [3] proposed a novel theory to explain the emotional experiences. He proposed to label emotional experience with categories such as love, joy, sadness, etc. Contrary to universal emotions proposed by Ekman [9], Semantic space theory expands the set of emotional categories to over twenty emotions compared to only eight in the universal emotions. It also argues that these emotional categories are embedded in a high dimensional space and have continuous gradients between them, with some categories such as joy and admiration being closer to each other compared to joy and sadness. Cowen *et al.* [2, 4, 5, 6] showed via human experiments that the Semantic space theory is better at explaining emotional experiences of human subjects compared to universal emotions as well as Affective dimension theory.
Demszky et al. [7] collected GoEmotions dataset which have reddit comments labeled according to the emotional categories of the Semantic space theory. In order to show that our $Combined$ dataset result in a better distribution over the emotional categories, we pretrained a BERT [8] model from HuggingFace [10] on GoEmotions and then used the pretrained model to predict the full set of emotional categories present in $Combined$ and $ArtEmis$. To highlight the quality of labeling in $Combined$ we calculate Pearson's correlation coefficient between all pairs of emotional categories. We present the results in Fig. 1. The histogram shows how $Combined$ is more balanced especially for the negative emotions such as fear, disgust, and sadness. The heat maps show that $Combined$ have less correlation between emotions which is desirable since it reflects the ability of the dataset to

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

better represent each emotion and make it distinct from other emotions. This effect is noticeable for the negative emotions since $ArtEmis$ had a bias towards positive emotions. For example, The row corresponding to Fear is much darker in $Combined$ with very dark patches compared the same row in $ArtEmis$ reflecting a better representation of the fear emotion in $Combined$. Surprisingly, the correlation between fear and other negative emotions such as disappointment went down drastically reflecting that $Combined$ captions better differentiate between negative emotions.

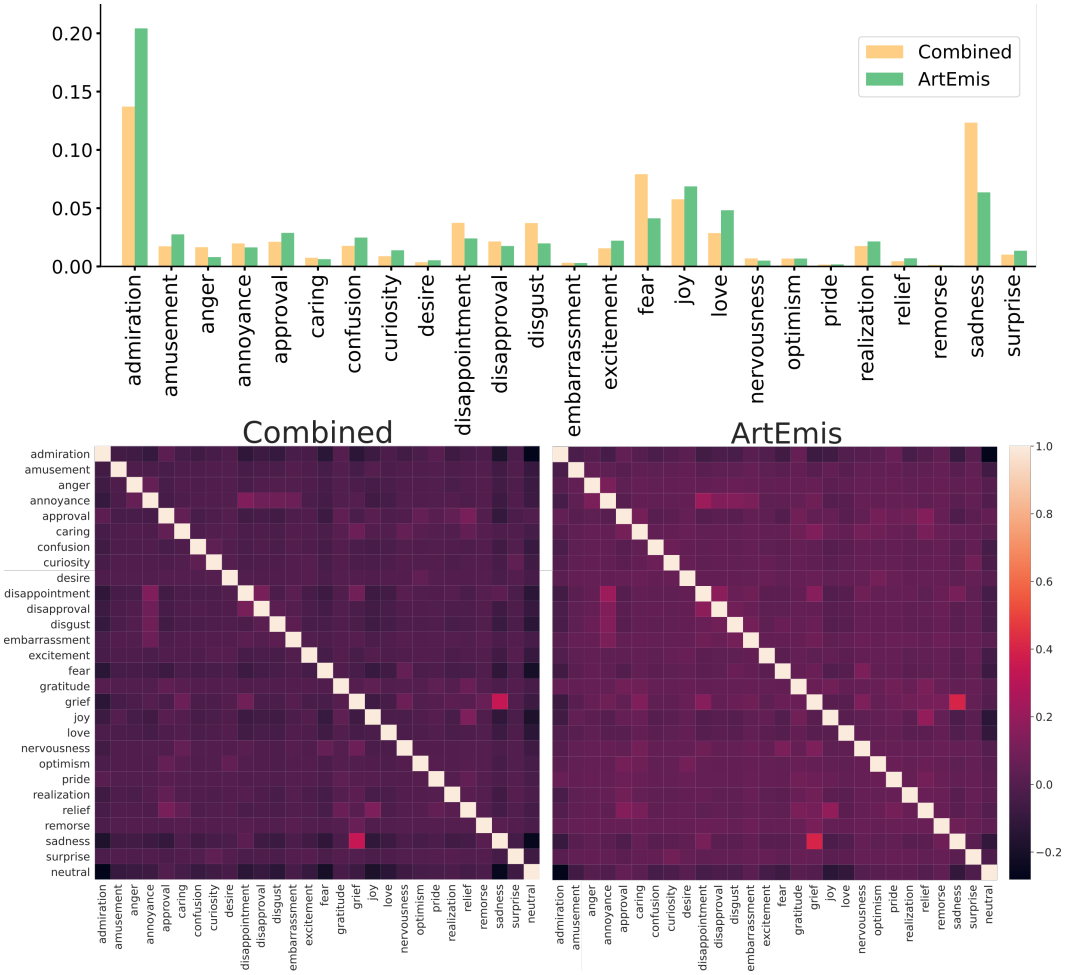


Figure 1: *Semantic Space Theory Fine-grained Emotion Analysis.* Top: we plot the histogram over the extended emotion set from GoEmotion. Note how the distribution is more balanced in $Combined$. Bottom: we show the correlation of the emotions in $Combined$ and $ArtEmis$. The darker off-diagonal patches mean $Combined$ has less correlation between different emotions, making it better at covering a wider range of emotional experiences.

Datasets		Metrics						
Test Set	Train Set	BLEU-0	BLEU-1	BLEU-2	BLEU-3	CIDER	METEOR	ROUGE
ALL	Combined	**0.628**	**0.385**	**0.226**	**0.137**	**0.103**	**0.165**	**0.339**
	ArtEmis	0.611	0.366	0.210	0.121	0.096	0.158	0.329
	Contrastive	0.616	0.367	0.211	0.125	0.096	0.160	0.439
	$\text{ArtEmis}_{0.5}$	0.612	0.366	0.209	0.120	0.093	0.157	0.320
C40	Combined	**0.860**	**0.672**	**0.483**	**0.337**	**0.094**	**0.220**	**0.454**
	ArtEmis	0.855	0.658	0.466	0.317	0.088	0.213	0.445
	Contrastive	0.818	0.610	0.422	0.284	0.082	0.204	0.429
	$\text{ArtEmis}_{0.5}$	0.851	0.654	0.457	0.304	0.084	0.212	0.443
NEW	Combined	**0.558**	**0.323**	**0.184**	**0.111**	**0.121**	**0.156**	**0.320**
	ArtEmis	0.529	0.295	0.161	0.091	0.103	0.146	0.305
	Contrastive	0.561	0.319	0.180	0.106	0.119	0.154	0.317
	$\text{ArtEmis}_{0.5}$	0.533	0.297	0.162	0.091	0.101	0.144	0.307
OLD	Combined	**0.547**	**0.303**	**0.165**	**0.095**	**0.091**	**0.143**	**0.300**
	ArtEmis	0.541	0.299	0.162	0.091	0.096	0.143	0.298
	Contrastive	0.532	0.283	0.147	0.082	0.079	0.136	0.289
	$\text{ArtEmis}_{0.5}$	0.544	0.300	0.161	0.090	0.094	0.142	0.301

Table 1: **Extended evaluation of SAT.** We report all the experiments done with SAT on all the datasets. Note how models trained on $Combined$ outperform all the other models.

2 Extended Neural Speaker Evaluation

Datasets We name our complementary data as $Contrastive$ and the original dataset from Achlioptas et al. [1] as $ArtEmis$. We combine both datasets to get $Combined$ and for ablations we introduce a subset of ArtEmis named $ArtEmis_{0.5}$ which has the same size as $Contrastive$. We use these four datasets to train our neural speakers. Our test sets are sampled from $Contrastive$ and $ArtEmis$. We constraint our test sets to have unique images not found in the training sets. We define 5 test sets: NEW sampled only from $Contrastive$, OLD sampled only from $ArtEmis$, ALL which is the concatenation of OLD and NEW, finally, we use $ArtEmis_{C40}$ which was introduced in [1] which has more than 40 captions per image. We guarantee that the number of captions per image in NEW and OLD to be more than five and less than 10 captions.

Extended Results We train two sets of SAT models on each training set and evaluate it on all the test sets. The first set is not emotionally grounded i.e. Vanilla SAT. The second set has emotionally grounded models which take as an input the emotional label and ground the decoder of the SAT on the input emotion. We report the results in Table 2. Note how $Combined$ is outperforming all the other training set on all of the test sets reflecting the effect of balanced emotion distribution and captions' distinctiveness.

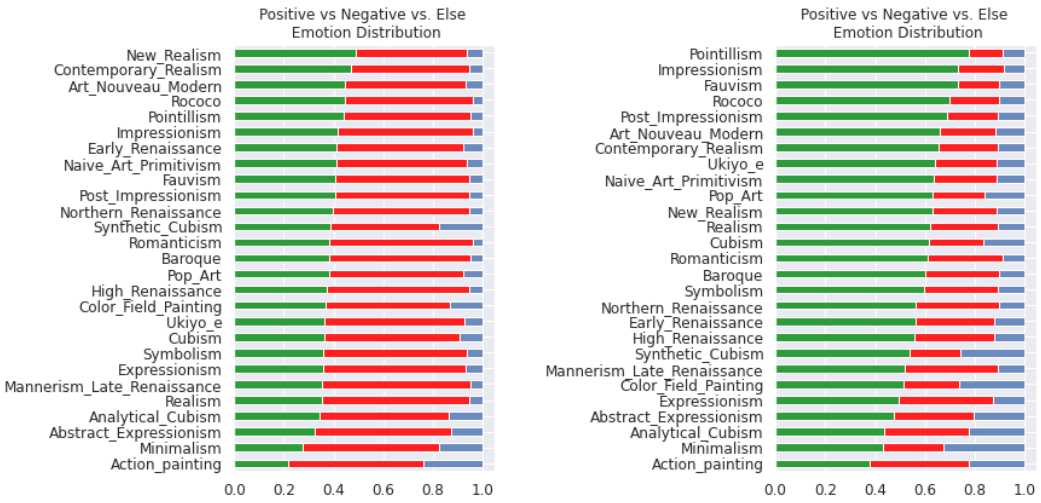
Datasets		Metrics						
Test Set	Train Set	BLEU-0	BLEU-1	BLEU-2	BLEU-3	CIDER	METEOR	ROUGE
ALL	Combined	**0.626**	**0.380**	**0.223**	**0.136**	**0.101**	**0.165**	**0.339**
	ArtEmis	0.609	0.367	0.209	0.120	0.095	0.159	0.330
	Contrastive	0.615	0.367	0.212	0.126	0.098	0.160	0.333
	ArtEmis$_{0.5}$	0.603	0.358	0.204	0.118	0.091	0.155	0.327
C40	Combined	**0.857**	**0.668**	**0.482**	**0.338**	**0.094**	**0.218**	**0.451**
	ArtEmis	0.841	0.646	0.458	0.313	0.084	0.213	0.446
	Contrastive	0.817	0.610	0.423	0.286	0.084	0.204	0.432
	ArtEmis$_{0.5}$	0.845	0.648	0.455	0.307	0.085	0.208	0.440
NEW	Combined	**0.558**	**0.320**	**0.183**	**0.110**	**0.121**	**0.156**	**0.319**
	ArtEmis	0.530	0.297	0.163	0.092	0.102	0.147	0.305
	Contrastive	0.558	0.318	0.180	0.105	0.122	0.154	0.318
	ArtEmis$_{0.5}$	0.523	0.288	0.156	0.088	0.097	0.142	0.301
OLD	Combined	**0.543**	**0.298**	**0.162**	**0.094**	**0.089**	**0.142**	**0.301**
	ArtEmis	0.539	0.298	0.159	0.089	0.096	0.143	0.300
	Contrastive	0.533	0.283	0.149	0.085	0.081	0.136	0.291
	ArtEmis$_{0.5}$	0.535	0.292	0.158	0.089	0.094	0.140	0.298

Table 2: **Extended evaluation of emo-grounded SAT.** We report all the experiments done with emo-grounded SAT on all the datasets. Similar to the non-grounded case, models trained on $Combined$ outperform all the other models.

3 Extended Linguistic Analysis

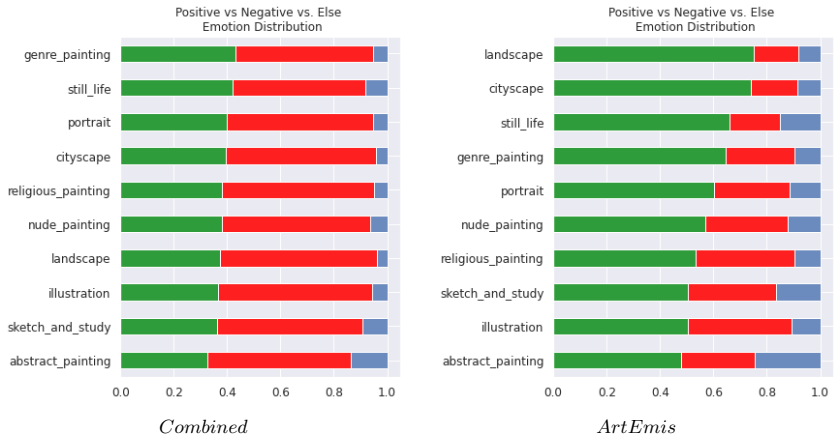
We provide in this section an in-depth linguistic analysis to highlight the differences between the contrastively collected data and the original ArtEmis data. Figures 2 and 3 compare the emotional sentiment distribution over different genres and art styles. We compare $ArtEmis$ and $Combined$ instead of $Contrastive$ because $Contrastive$ is collected as a complementary unbiasing dataset to the original ArtEmis dataset. $Combined$ dataset show clearly the balancing effect of contrastive data collection. This balancing is notable over most of the art styles and genres. The number of negative emotions increased because ArtEmis had 34779 Artworks with single sentiment emotions. Out of them, 29879 had a positive sentiment. Our contrastive method collected negative samples for these positive Artworks leading to a more balanced combined dataset.

In Fig 4, we report the concreteness, subjectivity and sentiment of the utterances in $Combined$ dataset. The values are very similar to those reported in Achlioptas et al. [1] for ArtEmis dataset, supporting the neutral effect of contrastive data collection on the utterance structure. Similar to ArtEmis, the reported subjectivity and sentiment is higher than average captioning datasets while the concreteness is lower, reflecting the nature of opinions on Artworks. This also shows that the improvements introduced by the newly collected data result from balancing the emotions and enhancing the quality of utterances by making them attend to fine details in the artworks.



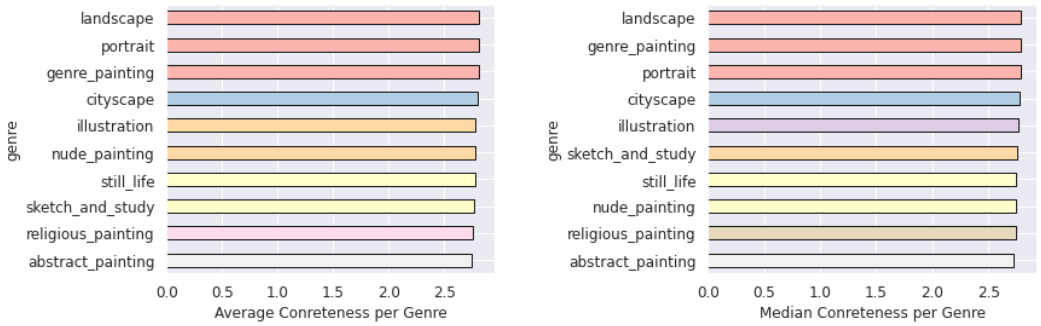
Combined *ArtEmis*

Figure 2: Emotion sentiment distribution per style

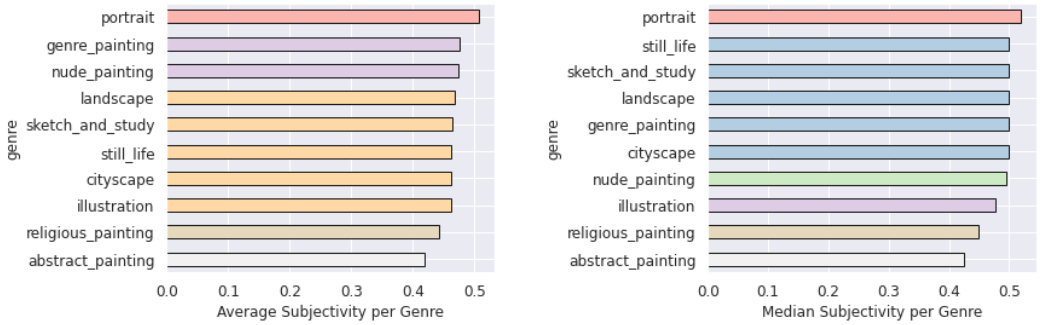


Combined *ArtEmis*

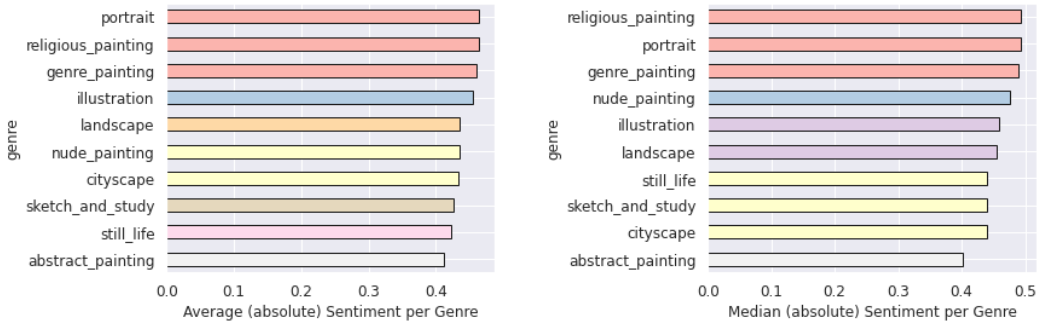
Figure 3: Emotion sentiment distribution per genre



(a) Average/Median Utterance Concreteness per genre



(b) Average/Median Utterance Subjectivity per genre



(c) Average/Median Utterance Sentiment per genre

Figure 4: Different linguistic measures in $Combined$

4 Neural Speakers Qualitative Analysis

We show in Fig. 5 examples of utterances generated conditionally on different emotions using a Show, Attend, and Tell neural speaker [11]. These utterances show the ability of the neural speaker to understand emotions and reflect them correctly. In all of the utterances, we can see an association between emotions and words. This is evident in the sample (e), where every emotion produces a different describing word that reflects the grounding emotion. Remarkably, example (b) has the word sky, which is not present in the painting, showing the imaginative aspect of affective datasets. Finally, example (f) reveals that different emotions attend to different objects and aspects. For example, anger is associated with messy colors, while amusement is more related to people having fun in the painting. This attention to detail is emphasized in the contrastive data collection leading to such distinct emotion-object relations. Figure 6 shows another set of unconditionally generated utterances. These utterances are sampled randomly without any emotional conditioning. These generated utterances show attention to details and a deeper understanding of the objects in the paintings, even for abstract paintings, such as the one in the top left.

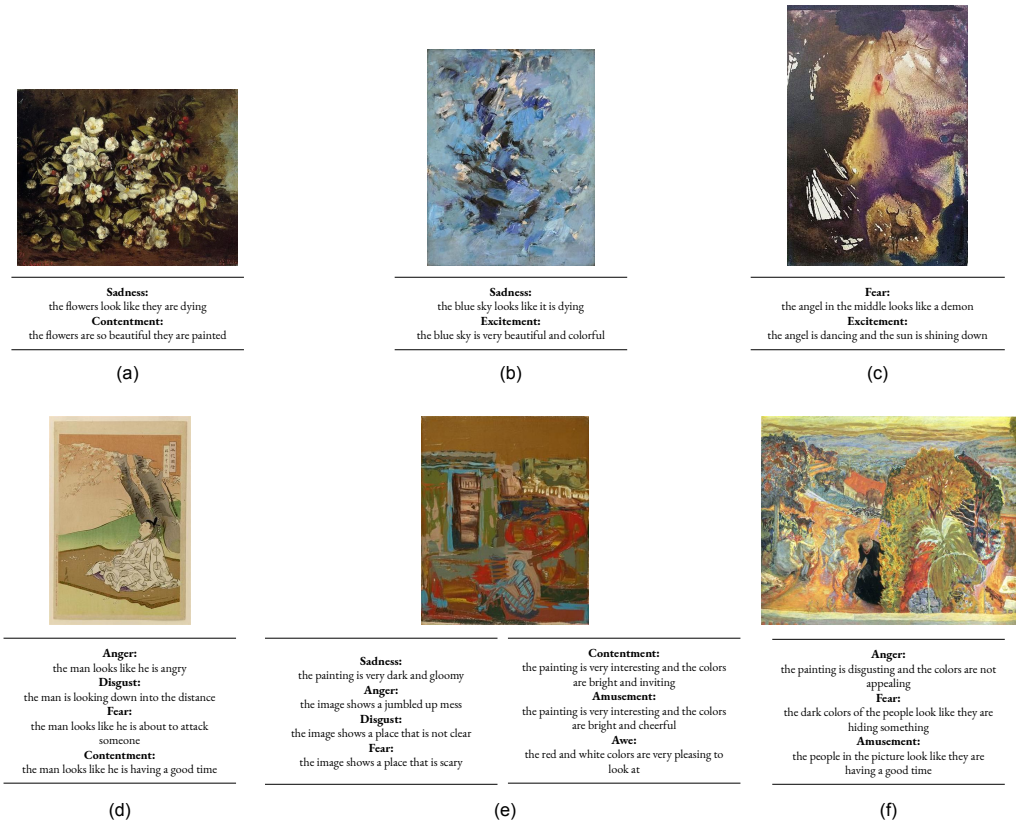


Figure 5: Examples of utterances conditioned on emotions generated using SAT model.

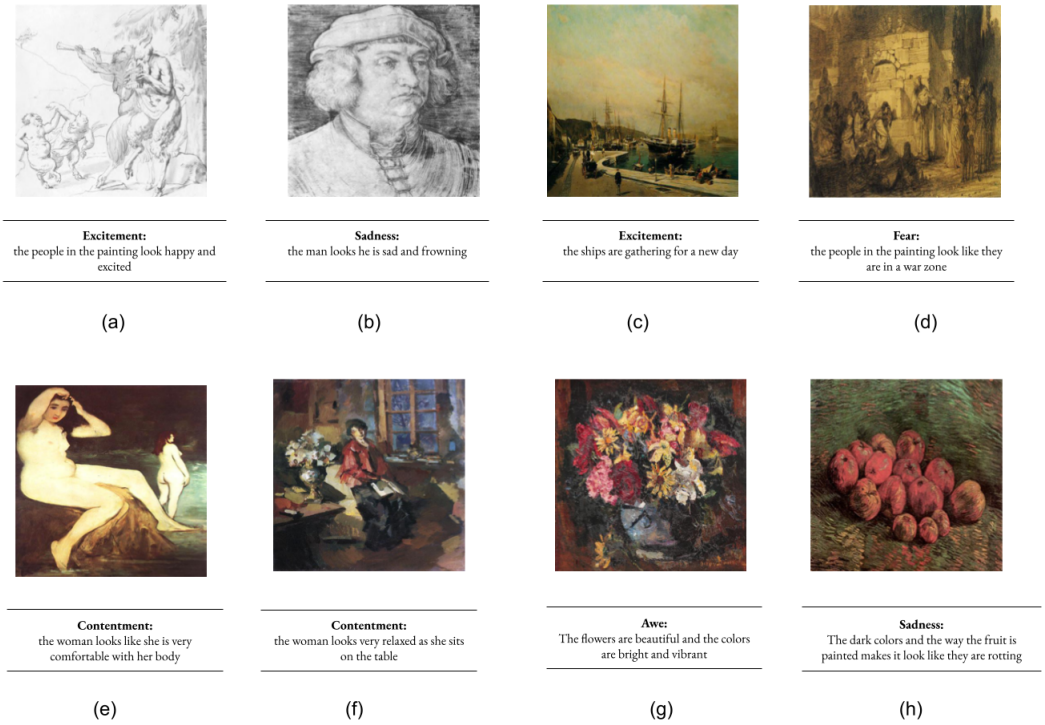
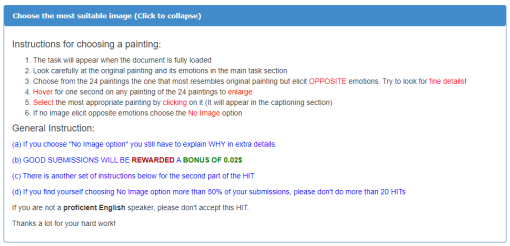
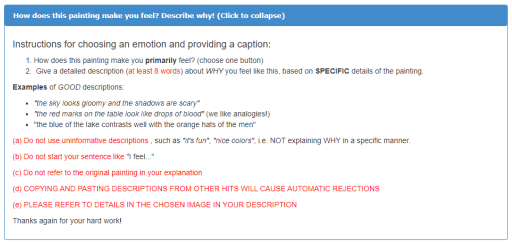


Figure 6: Examples of utterances generated unconditionally using SAT model.

Figure 7: Examples from the contrastively collected dataset.On the left side of each example is the query painting. The right side shows a similar painting, based on the VGG feature map, which evokes the opposite emotion. We show the old utterance of the selected image and the new utterance to highlight the increased attention to details. Despite of paired paintings having very similar styles, the triggered emotions and utterances are very different.



(a) Instructions for selecting the most similar painting and accepting the task



(b) Instructions for selecting an emotion and writing the utterance

Figure 9: The two sets of instructions used in Contrastive Data Collection

5 Dataset Statistics

Our newly collected *Contrastive* dataset can found at `www.artemisdataset-v2.org`. We used two sets of instructions for the Amazon Mechanical Turk workers. The first set, shown in Fig. 9a, contains general instructions about the task as well as instructions for selecting the most similar painting. The second set, shown in Fig. 9b, contains instructions for choosing the emotions and writing the utterances. The instructions explain the interface for the users and how they can interact with it. It also encourages users to submit quality work through bonus rewards. Finally, we provided a measure for the users to assess their suitability for the task. We ask them to stop participating if they submit many tasks with no applicable option. In the emotion selection and explanation instructions, we provide examples of phrases to avoid using since they lead to bad submissions (e.g., "it is fun", "nice colors"). Fig. 8 shows the frequency of the number of submissions per user. It can be seen that most users submit less than 20 submissions which add to the diversity of our collected dataset. Furthermore, the average time taken by a user to submit a task is 2.5 minutes, and we paid 10 cents per task resulting in an hourly rate of 2.4 dollars which is more than the average earning rate of Amazon Mechanical Turk (less than 2 dollars).

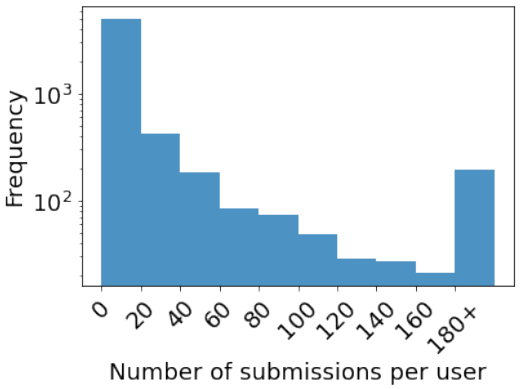


Figure 8: The frequency of submissions per user

We show more qualitative sample in Fig. 7 to compare the quality of *Combined* and *Contrastive*.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. In *CVPR*, 2021.

[2] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38): E7900–E7909, 2017.

[3] Alan S Cowen and Dacher Keltner. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, 2020.

[4] Alan S Cowen, Hillary Anger Elfenbein, Petri Laukka, and Dacher Keltner. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698, 2019.

[5] Alan S Cowen, Xia Fang, Disa Sauter, and Dacher Keltner. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, 117(4):1924–1934, 2020.

[6] Alan S Cowen, Dacher Keltner, Florian Schroff, Brendan Jou, Hartwig Adam, and Gautam Prasad. Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841): 251–257, 2021.

[7] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[11] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.