ArtEmis: Affective Language for Visual Art
Supplemental Material

Panos Achlioptas[1]
panos@cs.stanford.edu

Maks Ovsjanikov[2]
maks@lix.polytechnique.fr

Kilichbek Haydarov[3]
kilichbek.haydarov@kaust.edu.sa

Mohamed Elhoseiny[3,1]
mohamed.elhoseiny@kaust.edu.sa

Leonidas Guibas[1]
guibas@cs.stanford.edu

[1]Stanford University
[2]LIX, Ecole Polytechnique, IP Paris
[3]King Abdullah University of Science and Technology (KAUST)

A. ArtEmis building details

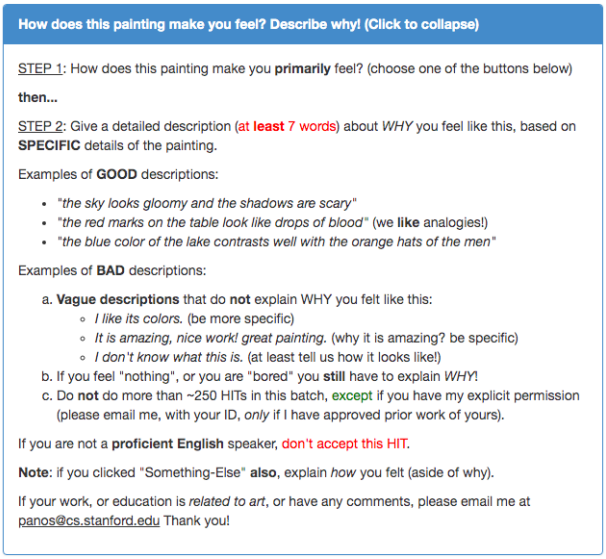


Figure 1. **AMT instructions for ArtEmis data collection.**

In total, we annotated 80,031 artworks covering the entire WikiArt, as downloaded in 2015 [6]. We note that this version of the WikiArt dataset contains 81,446 artworks. However, as our analysis indicated 1,415 artworks were exact duplicates, of the 80,031 *unique* artworks we kept for annotation purposes. We found these duplicates using the 'fdupes' program [5] and limited manual inspection on pairs of nearest-neighbors artworks (using features of a ResNet-32, pretrained on ImageNet), whose distance was smaller than a manually selected threshold.

When displaying the image of an artwork in AMT we scale down its largest size to 600 pixels, keeping the original aspect-ratio (or do not apply any scaling if the largest size is less than 600 pixels). We do this scaling to homogenize the presentation of our visual stimuli, and crucially to also reduce the loading and scrolling time required with higher resolution images.

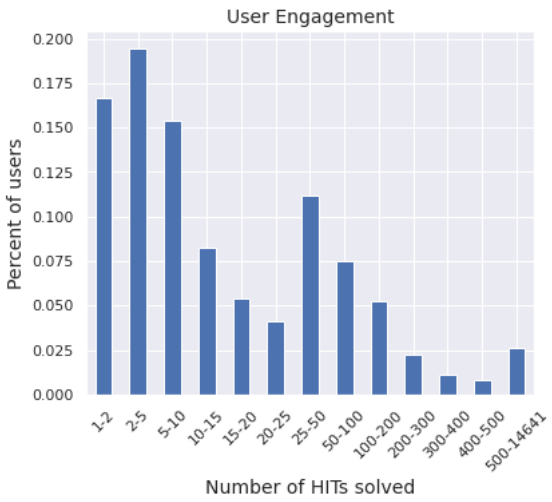


Figure 2. **User engagement in building ArtEmis.** The median and average number of annotations (HITs) solved by the AMT users is 9 and 67 respectively.

B. ArtEmis further analysis

Richness & Diversity To determine the quantities shown in Tables 1 and 2 of the Main paper we use the NLTK part-of-speech tagger [1].

Sentiment analysis. In addition to being rich and diverse, and as we might expect, ArtEmis also contains language that is sentimental. To demonstrate this we used a rule-based sentiment analyzer (VADER [4]) and measured the degree to which an utterance of ArtEmis carries positive, negative and neutral sentiment. Specifically, VADER estimates the valence for each of these sentiment states via a normalized scalar: 0 (least positive), to 1 (most positive). Furthermore, we followed the standard practice of computing a compounding metric to aggregate the three sentiment scores into a single scalar and, through an appropriate threshold, classify an utterance into one of the three sentiment types. By doing this, we found out that ArtEmis is more sentimental than many captioning datasets by a large margin. For example, this classifier assigns only 16.5% of ArtEmis to the neutral sentiment, while for COCO-captions it assigns 77.4%. Similarly, a random utterance of ArtEmis has a compound sentiment score (absolute value) of 0.44 while for COCO this score is 0.07 (p-val significant, see also Main paper Fig.3 (c)).

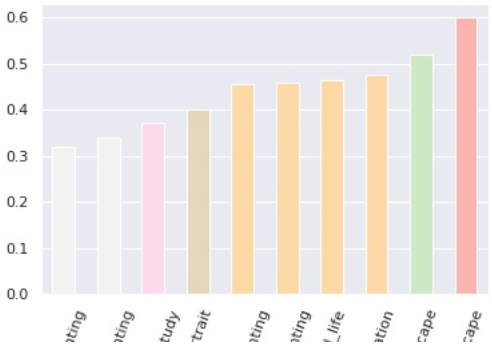


Figure 3. **User majority agreement in emotion, per genre**. Shown are the percentages of the artworks belonging in each genre for which the majority of the annotators chose the same emotion (or the something-else option).

Emotion-centric analysis, per genre. The genre of artwork where annotators achieve strong agreement most frequently is *landscape* paintings (60.0% of all such paintings), which is also the genre with most positive associated emotions (75.0% of the time). On the opposite end of the spectrum, *nude-paintings* achieve least frequently a majority: only 32%, while abstract artwork is the genre where the something-else category is selected most frequently (24.6%) and the one where the empirical emotion distributions on average (per painting, across annotators) have the largest entropy. We note that positive and mixed emotional reactions for landscapes and nude-paintings have been consistently observed in previous studies [3, 2] – see [3] for an interesting evolution-based perspective on the former phenomenon.

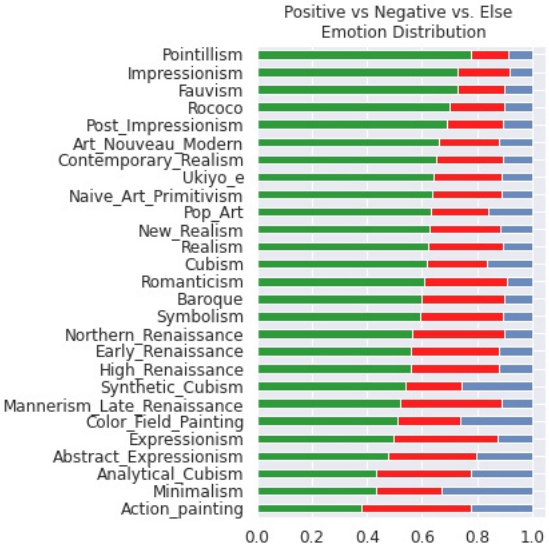


Figure 4. **Ternary user-based emotion distributions per artistic style.** The horizontal bars of each artistic style indicate the fraction of positive (color green) vs. negative (color red) vs. something-else (color blue) responses its artworks accumulated in ArtEmis. Each bar is scaled to 1. The styles are sorted in decreasing order of their positive fractions. More details are given in Paragraph B.

In Figures 4,5 we show a similar emotion-oriented analysis using the 27 artistic style annotations provided in [6]. Similarly to the previous analysis we first map the user indicated emotion to a positive vs. negative (or something-else) category. We show the resulting fractions of each category per art-style in Fig. 4. *Pointillism* is the style that has the largest fraction of its annotations being associated with positive emotions. This art style has also the lowest average entropy w.r.t. these three categories (Fig. 5).

Reasonableness user study. To assess if an ArtEmis utterance was a realistic and an emotionally fitting response to a given artwork, we ran a separate AMT user study. Specifically, we presented to users a total of 200 ArtEmis utterances with their corresponding artwork, and ask them to choose among four relevant options (see Fig. 6). Each artwork/utterance pair was inspected by 5 users. We aggregate their opinions, by associating with each annotation pair the option chosen most frequently among the users. The results, presented in the pie chart of Figure 7, reveal that the users consider the vast majority of the collected utterances (97.5% of all) realistic and emotionally reasonable responses to the underlying images. Interestingly, 51% of the annotations are marked as reasonable, but with the users stating that they would have reacted differently to the corresponding image. This last finding, further highlights the subjective character of our task.

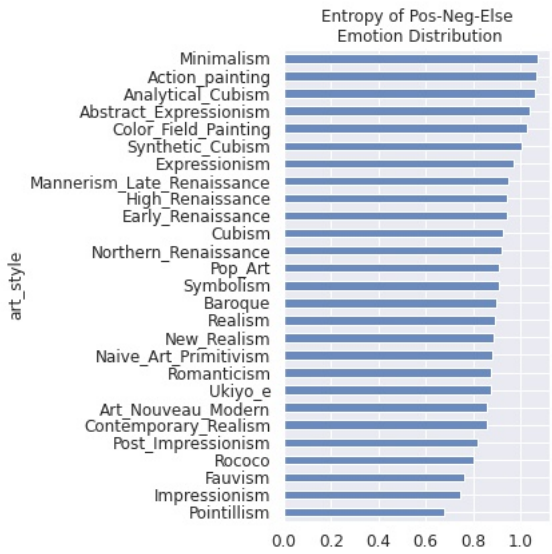


Figure 5. **Average entropy of emotion distributions, per artistic style.** For each annotated artwork we extract a ternary distribution based on the emotional responses of its (at least 5) annotators. The support of these distributions includes the positive, negative and something-else emotion categories. We compute the entropy of each derived empirical distribution and report here the average across different artistic styles.

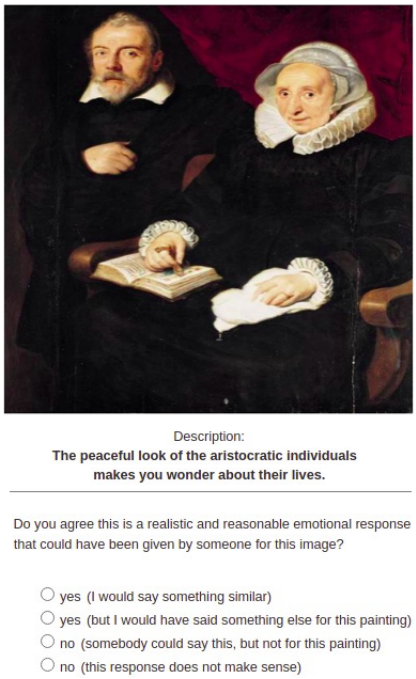


Figure 6. **Reasonableness AMT interface.** The users where given the options to strongly or weakly approve or disapprove the fitness of the caption to the painting.

Something-else option. We manually tagged the utterances explaining the something-else option to approximately find which are the emotions raised in this category. From the 52,962 utterances of this category, 5,333 include a word suggesting *confusion* (e.g., the words puzzled, or perplexed), 3,904 a word suggesting *boredom*, and 3,889 words suggesting *curiosity*.

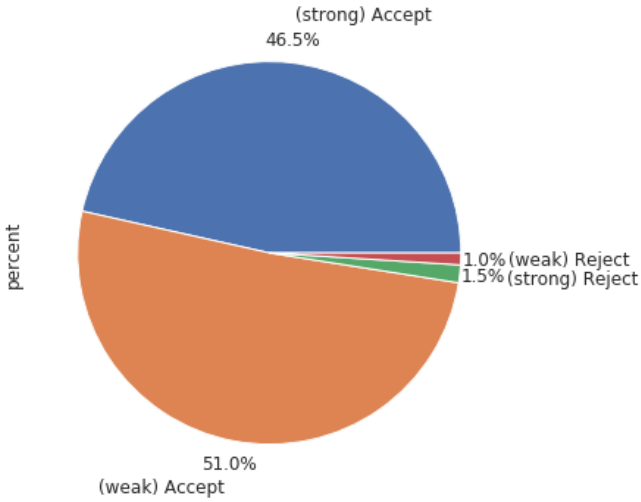


Figure 7. **Reasonableness Test.** The pie chart shows the proportion of utterances that fall in each of the categories the reflect different degrees of reasonableness.

C. ArtEmis miscellaneous

Figure 8. **Examples of unanimous positive reactions.** Characteristic examples where *all* annotators selected the same positive emotion (here, *contentment*). Users where significantly more likely to respond in a positive way to open landscapes, and colorful images depicting idyllic natural scenes.

D. Objective language for art

In order to deploy the ANP-speaker baseline described in Section 4.2 of the Main paper, we had to address first the domain gap between the typical images of the COCO-

Figure 9. **Examples of unanimous negative reactions.** Characteristic examples where *all* annotators selected the same negative emotion (here, *sadness*). Users where significantly more likely to respond in a negative way to dark colored images, with themes reflecting for instance, death or pain.

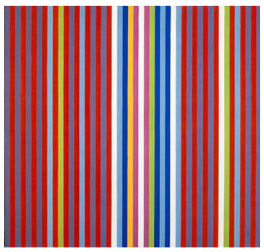
Figure 10. Wordcloud of the common words of ArtEmis. The size of each word is proportionate to its frequency.

captions and WikiArt. To this end, we collected a moderate size dataset, annotating 5,000 artworks of WikiArt, each with a single objective utterance describing the main items, parts, etc. found in artwork (See Fig. 11 for examples). Two exemplars of the effect that fine-tuning a pre-trained neural-speaker on COCO (SAT model) with this new dataset (dubbed ***OLA***, for Objective Language for Art) are shown in Figure 12.

E. Neural Net Studies

Failure neural-speaking cases. While the generations of the neural-speakers are intriguing in many cases and can even be thought as been made by humans (see Turing test in Section 6 of Main paper); they still have a long way to go to before they become as 'soulful' and diverse as their human-counterparts. The neural speakers are significantly less diverse than humans and can even make mistakes at the basic object-recognition level of reasoning, as shown in the examples of Figure 13.

"A woman in blue dress puts down her umbrella to hold some flowers."



"A lot of skinny vertical stripes are equidistant from each other."

"People are walking along a hilly street with buildings."

"A white house with some trees and bushes around the front door."

Figure 11. **Examples of objective descriptions for WikiArt paintings.** The captions are shown under each painting.

COCO: "*a statue of a man holding a skateboard*"

COCO + OLA: "*a woman in a red dress is sitting on a chair*"

COCO: "*a close up of a blue and white vase*"

COCO + OLA: "*a blue rectangle that is on a white background*"

Figure 12. **Effect of finetuning with OLA.** These are generations produced by a neural speaker trained only with COCO (**COCO**) vs. the same speaker further fine-tuned with OLA (**COCO + OLA**).

Effect of changing the grounding emotion. One of the benefits that the neural speaker that is grounded by emotion offers is the flexibility to steer its generations by a freely chosen, desired emotion. In Figure 16 we show some examples of using the distinct grounding emotions: those that correspond to the maximizer and the *second* maximizer image-to-emotion classifier described in Section 4.1 of the Main paper.

Qualitative comparison of various neural speakers. In Figure 14 we present sample generations of our various neu-

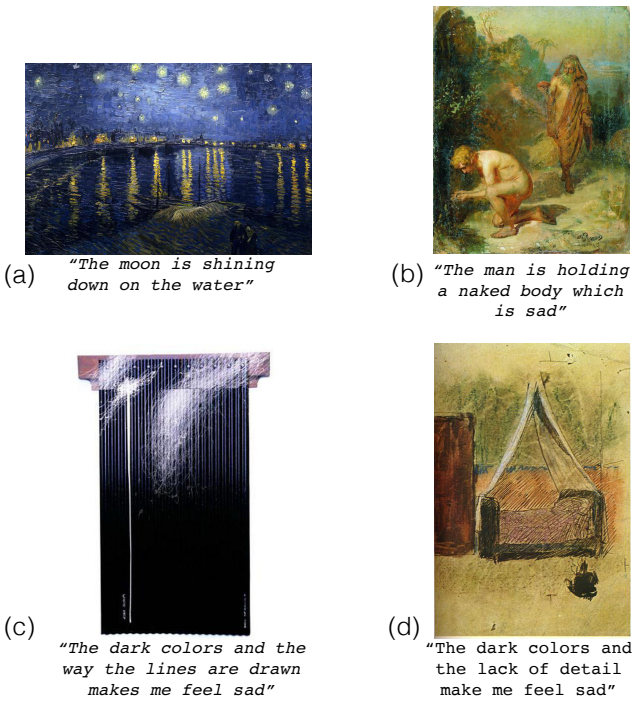


Figure 13. **Failing examples of neural generations.** The top-row examples capture wrongly the semantics: for (a) there is not a single moon, and (b) the man's body is naked but he is not holding it. The bottom-row examples exemplify how mode-collapse to 'vanilla' like (emotional) explanations can occur.

ral speakers.

Figure 14. **Test generations of different speakers.** The speakers models (indicated in boldfaced fonts) are those presented in the Main paper in Section 4.2.

References

[1] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009. 1

[2] Margaret Bradley, Maurizio Codispoti, Dean Sabatinelli, and Peter Lang. Emotion and motivation ii: Sex differences in picture processing. *Emotion*, 2001. 2

[3] Denis Dutton. *The Art Instinct: Beauty, Pleasure, and Human Evolution*. Bloomsbury Press, 2010. 2

[4] C.J. Hutto and Eric E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. eighth international conference on weblogs and social media. *ICWSM*, 2014. 2

[5] A. Lopez. *Fdupes is a program for identifying or deleting duplicate files residing within specified directories.*, (accessed 2020). Available at https://github.com/adrianlopezroche/fdupes. 1

[6] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *CoRR*, abs/1505.00855, 2015. 1, 2

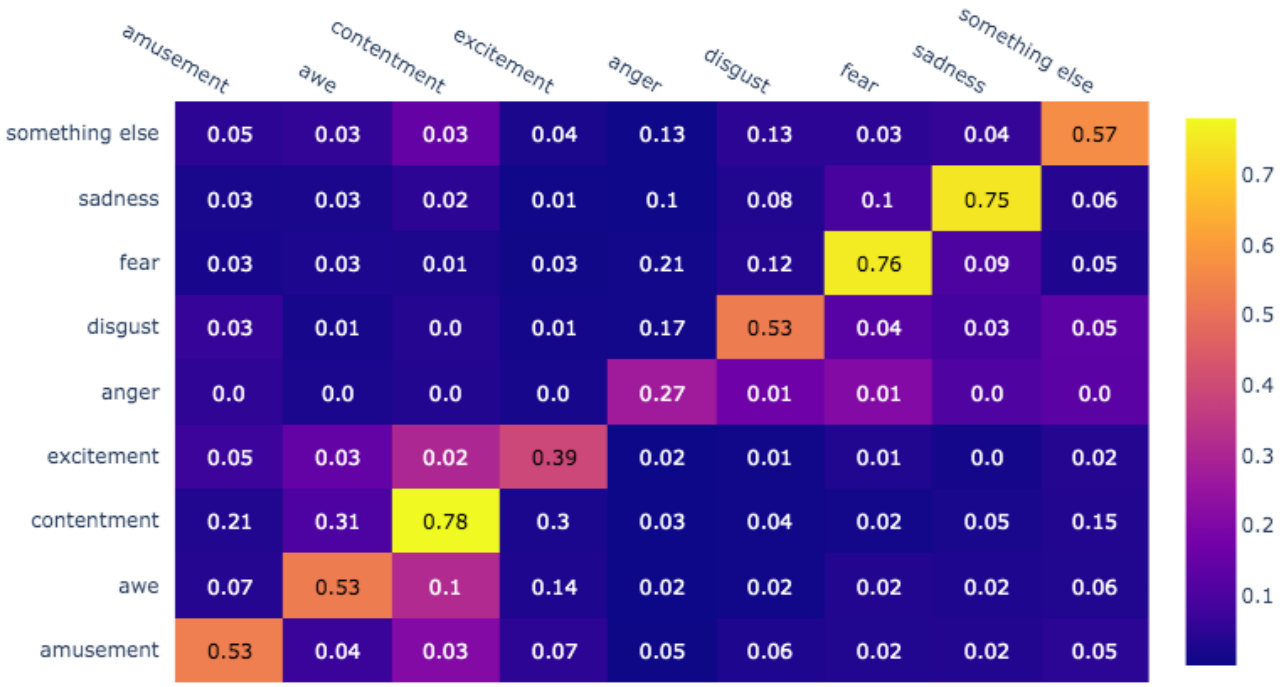
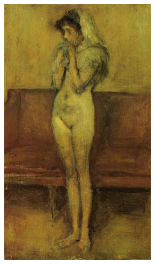


Figure 15. **Confusion matrix for text-only classification of emotion**. The results here are from the LSTM model described in Main, Section 4.1. Each column shows how percentage-wise the model confuses the specific emotion with all available emotion classes. Each columns sums to 1 (modulo rounding errors). The results are similar for a BERT text-classifier. Crucially, most confusion happens among emotions of the same-sentiment (positive, negative). Interestingly, the most misclassified class is that of anger, which is also the least frequently occurring class of ArtEmis.

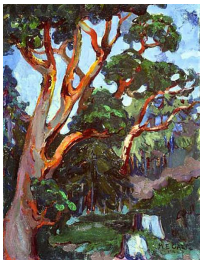


Sadness
"the woman looks like she is ashamed of her body"

Contentment
"the woman is nude and her body is very relaxed"

Contentment
"the women are enjoying their time together"

Sadness
"the woman looks like she is being forced to be in a hurry"



Fear
"the tree limbs look like they are screaming"

Contentment
"the colors are bright and vivid and the scene is very peaceful"

Awe
"the white mountains in the background are majestic and imposing"

Contentment
"the mountain peaks and the blue sky are very calming"

Contentment
"the man 's face is very calm and the colors are very neutral"

Awe
"the man 's fancy clothes and fancy uniform makes him look imposing"

Sadness
"the man looks like he is about to cry"

Contentment
"the man looks like he is thinking about something important"

Contentment
"the people in the painting look like they are enjoying a leisurely stroll"

Awe
"the detail in the architecture is amazing"

Something Else
"i am not sure what this is supposed to be"

Amusement
"the orange circle looks like a giant egg"

Figure 16. **Effect of changing the grounding emotion.** Shown are caption generations on test images with the SAT-speaker variant, based on a grounding input emotion (shown in bold above each caption). The grounding emotion with the highest (top) and second highest (bottom) scores are used as input. These emotions are inferred by a separately trained image-to-emotion classifier.