Affection: Learning Affective Explanations for Real-World Visual Data
Supplemental Material

Panos Achlioptas[1,3] Maks Ovsjanikov[2] Leonidas Guibas[3] Sergey Tulyakov[1]

[1]Snap Inc. [2]LIX, Ecole Polytechnique, IP Paris [3]Stanford University

A. Details on Building Affection

We build Affection by annotating images existing in the following five datasets: MS-COCO [4], Visual-Genome [15], Flickr30k Entities [21], Emotional-Machines [14], and the images considered in the work of Quanzeng *et al.* [22]. Specifically, we begin by annotating with affective responses *all* images in the latter two emotion-oriented works. We then proceed by using the images in Quanzeng *et al.* to find for each one of them its three nearest-neighbors in the image collections of MS-COCO, Visual-Genome and Flickr30k Entities, respectively. We include and annotate with affective responses the found neighbors, resulting in covering additionally 22,770 images from MS-COCO, 13,202 from Flickr30k Entities, and 16,437 from Visual-Genome.

To implement the nearest neighbor search we use the 512D embedding space formed by the output weights of the final convolutional layer of a ResNet-32 [9], pre-trained on ImageNet [7]. Before applying the search algorithm, we first average pool the 7×7 spatial dimensions of the penultimate ResNet layer (forming a $1 \times 1 \times 512$ embedding vector per image).

For the Visual Genome it is worth noting that we restrict the nearest-neighbor search on its 56,506 images (out of 108,077) that are *not* included in COCO [4], or Flickr30k Ent. [21], to enable the discovery of a larger number of *unique* neighbors across the individual datasets. As a final step, upon aggregating all relevant images from all corresponding (five) datasets, we use "fdups" [17] to remove possible duplicate images among them. For the final version of Affection, we detected and removed 198 duplicates found in this manner.

B. Analyzing Properties of Affection

In this Section, we briefly include some supplementary analysis, similar in spirit to the one presented in Section 3 of the main paper.

First, we count count the *unique* Parts-of-Speech (PoS) that different annotators use in their explanations for the *same* image (Table 1). We find that Affection has a significantly higher average number across all PoS than other datasets. This fact implies that our collected annotations are both lexically more rich (main paper, Table 1), and also more *diverse* than other datasets.

Dataset	Nouns	Pronouns	Adjectives	Adpositions	Verbs
Affection	**20.9** (3.4)	**4.4** (0.7)	**9.6** (1.5)	**8.6** (1.3)	**18.7** (3.0)
ArtEmis [2]	18.7 (3.4)	3.1 (0.6)	8.3 (1.5)	6.5 (1.2)	13.4 (2.4)
Flickr30k Ent. [32]	12.9 (2.6)	0.8 (0.2)	4.0 (0.8)	4.9 (1.0)	6.4 (1.3)
COCO [4]	10.8 (2.2)	0.6 (0.1)	3.3 (0.7)	4.5 (0.9)	4.5 (0.9)
Conceptual Capt. [25]	3.8 (3.8)	0.2 (0.2)	0.9 (0.9)	1.6 (1.6)	1.1 (1.1)
Google Refexp [19]	7.8 (2.2)	0.4 (0.1)	2.8 (0.8)	2.9 (0.8)	2.3 (0.6)

Table 1. **Lexical comparison over distinct part-of-speech categories, per *individual images***. The shown numbers indicate unique words per category averaged over individual images. In parentheses, we include a normalized version accounting for discrepancies in the number of annotators individual images might have. Evidently, Affection's language is lexically more **diverse**.

Second, we analyze how some of the key linguistic properties discussed in that main paper, are manifested in the annotations collected for each of the five underlying image datasets used to build Affection. Namely, we report the average attained scores (computed with the same methods described in the main paper) for the properties of *concreteness*, *subjectivity* and use of *sentimental* language. As seen in Figure 1, the annotations collected based on images found in the emotion-oriented datasets of Emotional-Machines [14] and of Quanzeng *et al.*, result on average in only *slightly* more abstract, subjective and sentimental affective explanations (language), compared to used images from the remaining datasets. In other words, it appears that w.r.t. these critical characteristics of Affection, the images used across *all* underlying datasets do not result in any significant discrepancy among the responses they evoked (hence, ameliorating concerns of a possible bias among them).

Third, for the above described properties, we also compare Affection to ArtEmis [2] (Figure 2). As mentioned in the main paper Affection and ArtEmis are similar in terms of their average concreteness scores (average scores of 2.82 vs. 2.81), but also Affection contains significantly more subjective and sentimental annotations (see histograms (b) and

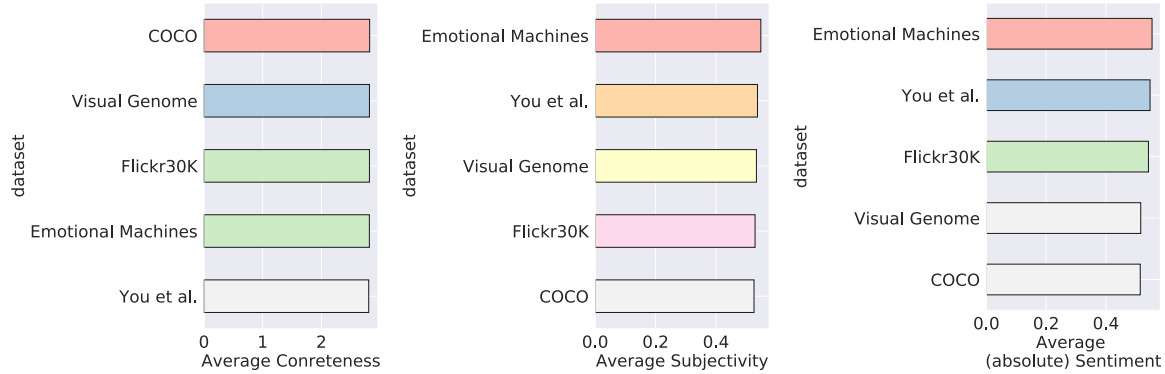

Figure 1. **Measuring key properties of Affection across its underlying image datasets**. Histograms comparing Affection in each of its underlying image datasets along the axes of (a) *Concreteness*, (b) *Subjectivity*, and (c) *Sentiment*.

(c) of Figure 2, Figure 7, and Section 3.1 of main paper).

Remark on Joint Data Exploitation. Affection is built on top of images for which rich annotations exist that are *complimentary* to our collected affect-oriented annotations. For instance, the images from Emotional-Machines [14] contain Valence-Arousal measurements (see first Paragraph of Section 2 in main paper), and the images from Quanzeng *et al.* contain further categorical image-to-emotion-classification labels. Most importantly, for FlickR30K, Visual Genome, and COCO, descriptive captions accompany **each** annotated image of Affection. We believe that a joint exploitation of those annotations with the data in Affection offers many promising future directions, e.g., one can imagine neural speakers that disentangle and control the 'objective' parts of our visually grounded explanations, from their more subjective/personal references.

C. Data preprocessing

For all experiments described in Section 6 of the main manuscript, we train neural networks by using an 85%-5%-10% train/val/test split of Affection, making sure that the splits have no overlap in terms of their underlying images. Moreover, we ignore explanations that contain more than 51 tokens (99-th percentile of token-length across Affection), or those containing fewer than 5 tokens (in total these two constraints remove $\sim 1\%$ of all utterances). Following common practice (e.g., [2, 5, 11, 20],), we convert our captions to lower-case, remove punctuation characters, and perform tokenization with the NLTK toolkit [3]. We note that tokens appearing less than twice in the training set were replaced with a special <unk> token denoting an out-of-vocabulary word.

D. Fine-grained Emotion Classification from a Single Modality

As stated in Section 6 of the main paper, all auxiliary emotion classifiers trained with Affection fail **gracefully**, as in, they primarily confuse fine-grained emotion classes of the same (positive or negative) sentiment. Here we include the corresponding confusion matrices for the ResNet101-based *image*-2-emotion classifier (Figure 3) and the LSTM-based *text*-2-emotion classifier (Figure 4).

For the text-2-emotion classifiers specifically, it is worth noting that if we binarize their output predictions for the original 9-way posed problem along with the ground-truth labels into positive vs. negative sentiments (ignoring the something-else category); the LSTM-based, and the transformer-BERT-based models, achieve **94.0%, 95.5%** accuracy, respectively.

Finally, we note that for the image-2-emotion classifier we use during training and inference, only images for which there is a strong majority among the annotators w.r.t. the emotions they indicated. Crucially, as stated in the main paper, the underlying distribution of emotions when considering only such images is highly imbalanced (see Figure 8).

E. Neural Comprehension of Affective Explanations

As mentioned in the first paragraph (Section 4) of the main paper, we explore the extent to which the textual explanations in Affection refer to discriminative visual elements of their underlying images, to enable their *identification* among arbitrary images.

For the corresponding CLIP-based experiments of Section 6 we use a pretrained CLIP model with 400M parameters (version `ViT-B/32`). During inference, we couple all ground-truth image-caption/explanation pairs of a dataset with a varying number of uniformly randomly chosen images from the same dataset – and upon embedding them in CLIP's joint visio-linguistic space we retrieve for each given

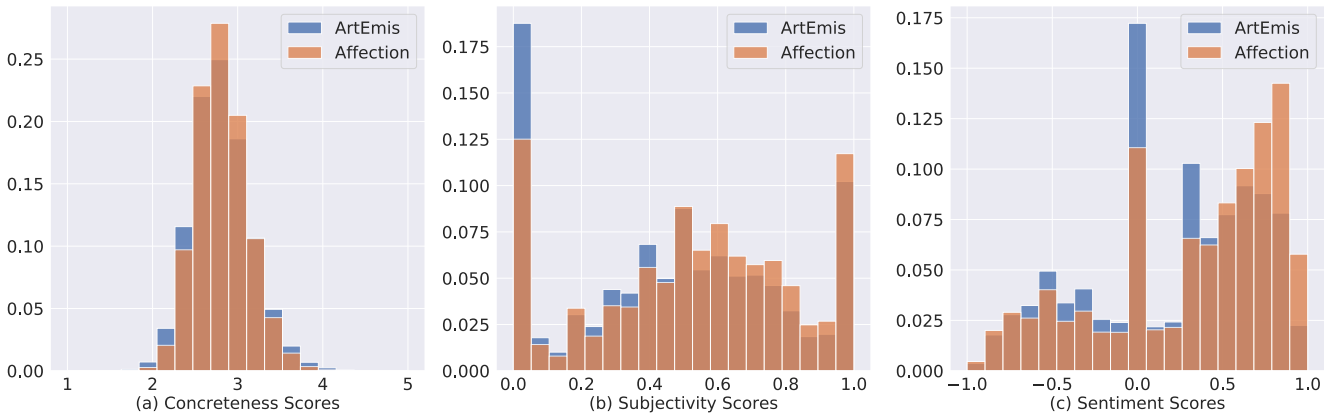

Figure 2. **Comparing Affection to ArtEmis** along the axes of (a) *Concreteness*, (b) *Subjectivity*, and (c) *Sentiment*. The histograms presented here are analogous to those contrasting Affection to COCO in Figure 2 of the main paper.

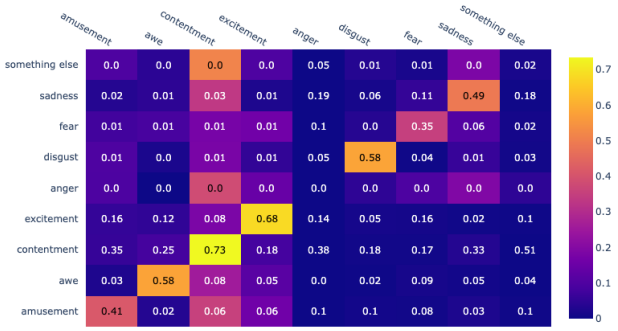

Figure 3. **Confusion matrix for a ResNet-101 pretrained image2emotion 9-way classifier trained and tested with Affection's emotion labels**. Only images and emotion labels for which there is a unique strong majority among the dominant emotions indicated by the annotators are used in this experiment.

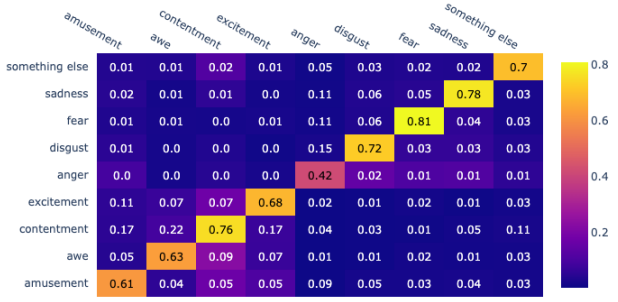

Figure 4. **Confusion matrix for an LSTM-based text2emotion 9-way classifier trained and tested with Affection's explanations**.

caption the image with the largest (cosine-based) similarity.

We note that to the best of our knowledge, these comprehension/listening studies are the first that address the extent to which *affective language is also referential* [27].

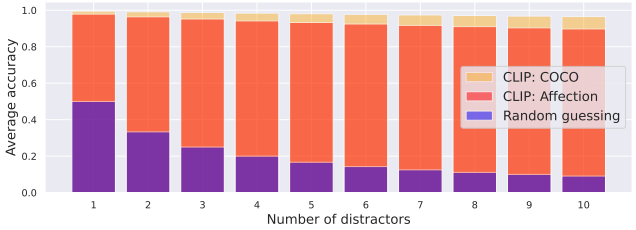

Figure 5. **Listening accuracy of a pretrained CLIP on the entire collection of Affection and COCO as a function of the number of distracting images used at inference time**. The x-axis displays the number of distracting images, and the y-axis the average accuracy of identifying the corresponding image given a ground-truth caption from either datasets. Random guessing reflects performance when selecting the target uniformly at random. Surprisingly, it appears that Affection contains explanations that describe salient visual elements regarding the image content, to enable *excellent* identification of them, i.e., comparably to the performance of using a purely objective dataset such as COCO.

As can be seen in the results of the average retrieval accuracy displayed in Fig. 5, Affection's explanations contain significant amounts of 'objective' and discriminative grounding details to enable **excellent** identification of an image from its underlying explanation. Specifically, the average accuracy when contrasting the ground-truth pair with a single distracting image is in the very high nineties for both datasets (COCO: 99.5% vs. Affection: 97.9%). Moreover, even with as many as ten distracting images the retrieval accuracy remains strong (COCO: 96.5% vs. Affection: 89.7%). Interestingly, for either dataset, the drop in performance with the addition of more distracting images is robust (less steep drop than guessing uniformly at random). Finally, we note that the training set of CLIP includes web-scale internet-crawled data, which are expected to be closer to COCO's nature than to Affection's i.e., affective explanations are possibly not as common online as descriptive image captions –

potentially explaining some of the performance gap observed among the two datasets.

Training from scratch. Figure 9 shows the test performance of a contrastive neural listener made by a transformer-based language encoder coupled with a ResNet-101-based image encoder; trained from scratch with Affection's explanations. Despite, the fact that this listener uses the explicit supervision of our annotations during training, it performs on average worse than the non-finetuned CLIP-based model [23] presented above. This fact is presumably due to using a combination of simpler architectural components (e.g., compared to the Vision Transformer [8] of CLIP), and not exploiting any web-scale pre-training.

F. Neural Speakers for AEC

Below we provide additional details for our default, emotion-grounded and pragmatic speaker variants mentioned in Section 4 of the main paper. It is important to note that our adaptations for the neural speakers (i.e., using emotion-grounding, and pragmatic-inference) are **generic and agnostic** to the choice of the base/backbone model.

Default speaker backbones. Our first backbone uses the Show-Attend-and-Tell (SAT) [31] backbone. The main element of this backbone is an LSTM cell [10] which is grounded with the input image and which by using Affection's data learns to generate utterances that explain plausible emotional reactions to it. Specifically, at each time step the model learns to attend [33] to different parts of the image (which is encoded by a separate ResNet-101 network), and by combining the 'current' input token with the LSTM's hidden state, attempts to predict the 'next' token. The output predicted token at each step is compared with the 'next' ground-truth token, under a cross-entropy loss using the paradigm of Teacher-Forcing [30]. To find a good set of model hyper-parameters (e.g. L_2-weights, dropout-rate and # of LSTM neurons) and the optimal (early) stopping epoch, we use a held-out validation set from Affection and select the model whose generations minimize the negative-log-likelihood against the ground-truth.

The second backbone uses the recent SoTA transformer-based architecture of GRIT [20]. Aside from replacing the LSTM cell and ResNet encoder of SAT with visual and language transformers, two noticeable differences of this backbone from SAT are that: i) it uses both region-based [6] and grid-based features [13, 18] (SAT only uses the latter), ii) it also uses self-critical training with CIDEr [29] score as the reward [24] besides cross-entropy-based Teacher-Forcing [30]. The effect of these changes is a significantly boosted attained score on all metrics capturing the similarity of the output generation to the held-out ground-truth explanations, compared to SAT as seen in Table 2 (vs. Table 2 of the main paper). However, when comparing these tables we also observe a noticeable deterioration of GRIT in terms of diversity of productions. E.g., a reduced number of unique productions, and higher Max-LCS and ClipDivCos. Despite these discrepancies, we also notice that the general trends and comparisons between their Default, Emo-Grounded and pragmatic variants are similar for the two backbones. E.g., the Pragmatic variants in both cases maximize the diversity and CLIP-Score-based metrics, and the Emo-Grounded variants maximize the Emotional-Alignment score.

It is important to note that for all shown qualitative neural speaking results and the results in Table 2 (and main paper, Table 2), our neural speakers are sensitive to the choices made during *inference* for i) the speaker's (soft-max) temperature, ii) the beam-size of the beam-search sampling, and iii) the relative importance we assign between the listening vs. speaking compatibility in the pragmatic variants. However, the trends these hyper-parameters create w.r.t. the machine-based evaluation metrics and specifically regarding the 'Best Strategy' (Table 2) one should follow to maximize each metric, are very stable and predictable [1, 2, 28].

Specifically, during inference for all the SAT-based neural speaking variants and the results presented in the main paper, we use beam-search with a beam size of 20 (or 5 for GRIT-based speakers) and a soft-max temperature for the layer predicting each generated token of 0.3. For the pragmatic variants of both backbones, the β parameter described in Section 4 (main paper) controlling the influence of the internal/judging listener is set to 0.25. When using GRIT as a backbone, for its Emo-Grounded variant, we use the public implementation of the original authors, by slightly adapting its updated class embedding $\mathrm{g}\langle\mathrm{cls}\rangle$ into an 9-dimensional (instead of 8-dimensional) vector, to include also the 'something-else' category, using a linear projection. During training, and following the authors' strategy we minimized the summation of two loss terms, i.e., for emotion prediction and caption generation.

Failure modes. Figure 6 displays examples of some of our neural speakers' characteristic (common) failure modes. The first problem oftentimes faced by *all* of our speaker variants is their inability to recognize the underlying object classes of the depicted objects in the grounding image. Thus, their generations might appear to ground their explanations on objects not actually displayed, e.g., describe properties of a male human when only females are shown. This generic error appears in numerous captioning systems and is not specific only to speakers trained with affective explanations. However, this problem can be more severe in typical affective imagery since such images tend to have more subtle and abstract semantics (e.g., pizza-like-looking wall clock, example (A)). A second but less frequently occurring problem that is also faced by all speaking variants is that they can sometimes create non-sensible emotional assessments, e.g.,

a human would find it strange to describe a bicycle as being calm (example (B)). Besides these generic problems, the main idiosyncratic problem we observed with the emotion-grounded variant is that it can overfocus (compared to other variants) on language concerning the underlying emotion while missing to ground key visual details. For instance, for image (C), the default variant produces '*I feel sad because the monkey looks like he is* **trapped in a cage**'. Finally, the pragmatic variant, unlike the emotion-grounded one, sometimes might try too hard to use specific visual details in its explanations, creating errors like those seen for image (D) – for which the default variant produces '*The zebras are beautiful and I would love to see them in the wild*'.

Emotional Turing test. As mentioned in Section 6 of the main paper, we evaluate how likely our neural speaking variants' output generations can be perceived as if they were made by humans. Specifically, we first form a random sample of 500 *test* images and accompany each image with one of their ground-truth, human-made explanations. We then couple each such image/explanation with a generation made by a neural speaker. We do this by considering all our four speaking variants to obtain 2,000 image-caption samples (an image paired with two explanations, a 'neural-' and a 'human-' based one). We then proceed by asking AMT annotators who have never seen these sampled images to observe them and, upon reading closely the coupled explanations, to select one among four options indicating that: (a) *both* explanations seem to have been made by humans justifying their emotional reaction to the shown image; (b) *none* of the explanations are likely to have been made by humans for that purpose, or (c) (and (d)) to select the explanation that seems more likely to have been made by a human. The findings of this emotional Turing test are summarized in Figure 10. As can seen in this figure, for all variants, more than 40% of the time (41.4%-46.2%), both displayed utterances were thought as if humans made them (blue bars). Moreover, and perhaps somewhat surprisingly at first reading, in a significant fraction of the answers, the neural-based generations were deemed more likely/fitting than the human-made ones (green bars of the same figure). These results highlight both the complexity of the AEC problem as well as the promising overall quality of our neural speaker solutions, enabled by the Affection dataset.

G. Ethical Considerations and Limitations

In this section we discuss certain considerations and limitations of the Affection dataset, along with presenting a short and final critique of methods utilizing it, including those we proposed by this work.

First, we highlight that our corpus contains emotional explanations given explicitly and only in English. Despite the fact that we collaborate with a large number of annotators (6,283) from different countries where English is the official spoken language (e.g., USA, Canada, etc.); its monolinguistic nature limits its universality making it more prone to possible cultural biases of the underlying ethnic groups. Second, sometimes images that are likely to induce an emotional reaction can portray sensitive topics including distressing situations such as injuries, riots, etc. – or explicit and sensitive topics such as nudes. Despite this fact, and importantly, we stress that for our studies: a) we use images already existing in public and widely adopted research datasets, b) we inform and ask for explicit consent of all our annotators to acknowledge the possibility of being exposed to such content, *before* allowing them to participate in our study. Specifically, our Amazon Mechanical Turk -based study requires each participant to have an 'Adult Content Qualification', does not collect any personally identifiable information for the participants, and strictly follows all guidelines described in the platform's acceptable use policy [26]. Last, we point out that an exemplar of possible misuse of Affection and methods learning from a dataset of this nature would be using it for finding or generating images that could induce a powerful emotion in an individual to manipulate them without their knowledge or consent. We do not endorse such usage and we ask for future users and researchers working with Affection to use it responsibly per the principles indicated in *Lo Piano* [16].

On the more technical side, AEC is a novel but subjective task, and poses new challenges for the learning community as mentioned in Sections 1 and 2 of the main paper. First, it requires matching not a single outcome but a distribution of outcomes (captions/explanations) to an underlying visual input. Second, it requires having metrics to capture how specifically an utterance's content is expressed in the aforementioned match. This is harder than typical caption quality assessment due to the increased subjectivity and corresponding variance of our underlying annotations; which makes reference-free-like metrics like CLIPScore or ClipDIVCos (Section 5 main paper) a promising starting direction. Third, we point out that there is not yet an established or easy way to capture how well an arbitrary explanation represents a justification for an emotion in a way that is informed by both visual and linguistic elements. Such a generic 'reasonableness'-like evaluation might require explicitly modeling intuition and other forms of logic and human knowledge that are intrinsically hard; making human-based evaluations like our Turing Test irreplaceable for the time being. Last, as we progress with automating and improving different aspects of the evaluation process, we expect the quality of the resulting neural-speaking models to also benefit from similarly improved loss functions. I.e., in the future, we expect to see neural models that incrementally approach human-level accuracy and diversity that avoid mistakes like those presented in Figure 6.

(A)
(Default speaker)
"the pizza looks so delicious and I would love to eat it"

(B)
(Default speaker)
"the bike is so calm and the trees are so beautiful"

(C)
(Emo-grounded speaker)
"this monkey looks sad and I feel sad for him" *

(D)
(Pragmatic speaker)
"the two zebras in the water look like they are having a good time"

Figure 6. **Most common failure modes of our affective neural speakers.** Left-most two examples show *generic* problems that *all* neural variants might suffer from: e.g., misidentifying the underlying visual elements (example A) or making non-sensible emotional judgments (example B). While the third example (C) is sensible, it highlights how an emo-grounded variant can overfocus on the underlying emotion and miss crucial visual details (e.g., the fence). On the contrary, the pragmatic variant (example D) can overcompensate by wrongly mentioning visual details (the default neural speaker simply mentions the zebras in this example). For more details see Section F

Metrics	**Default**	**Emo-Grounded**	*Speaker Variants* **Default (Pragmatic)**	**Emo-Grounded (Pragmatic)**	*Best Strategy*
BLEU-1 (↑)	**72.1**	71.3	72.0	70.9	default architecture
BLEU-2 (↑)	**41.7**	39.9	41.3	39.8	
BLEU-3 (↑)	**24.2**	22.7	23.9	22.9	
BLEU-4 (↑)	**14.5**	13.6	14.1	13.5	
METEOR (↑)	15.7	15.0	**15.9**	15.3	
ROUGE-L (↑)	32.3	31.4	**32.5**	31.9	
SPICE (↑)	7.9	7.2	**8.2**	7.6	
CLIPScore (↑)	67.6	67.7	**71.1**	**71.0**	pragmatic
RefCLIPScore (↑)	77.1	77.0	**78.2**	**78.2**	
Unique-Productions (↑)	76.5	78.8	81.1	**81.4**	pragmatic
Max-LCS (↓)	71.7	71.6	71.2	**70.4**	
ClipDivCos (↓)	74.2	73.9	70.6	**70.3**	
Similes (↓)	40.0	35.1	39.7	**33.8**	emo-grounded architecture
Emo-Alignment (↑)	50.1	56.8	50.3	**57.6**	

Table 2. **Neural speaker machine-based evaluations with GRIT-based backbone [20]**. The Default models use for grounding only the underlying image, while the Emo-Grounded variants also input an emotion-label. Pragmatic variants use CLIP to calibrate the score of sampled productions before selecting the final proposal.

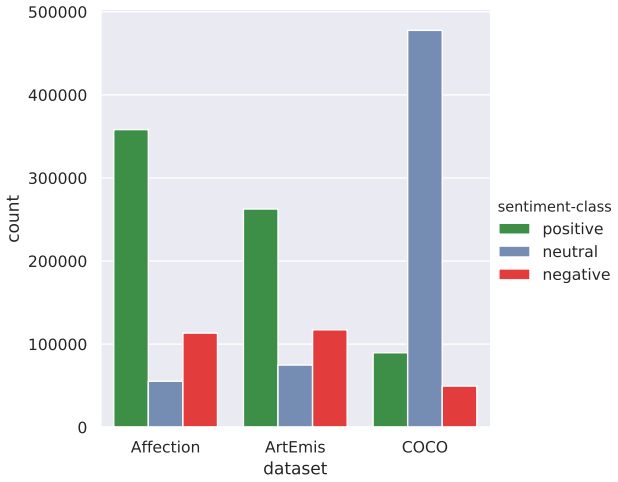

Figure 7. **Sentiment classes per dataset**. Using VADER's [12] sentiment classifier to assign the utterances of the shown datasets, in one of three classes. Affection's utterances are on average consider the least neutral.

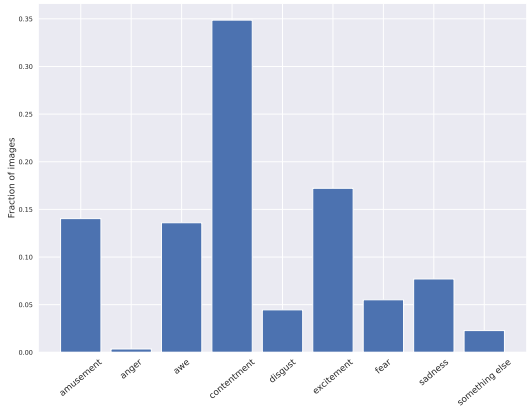

Figure 8. **Fraction of Affection images that have a unique strong majority w.r.t. the dominant emotions indicated by Affections' annotators, per each emotion class**.

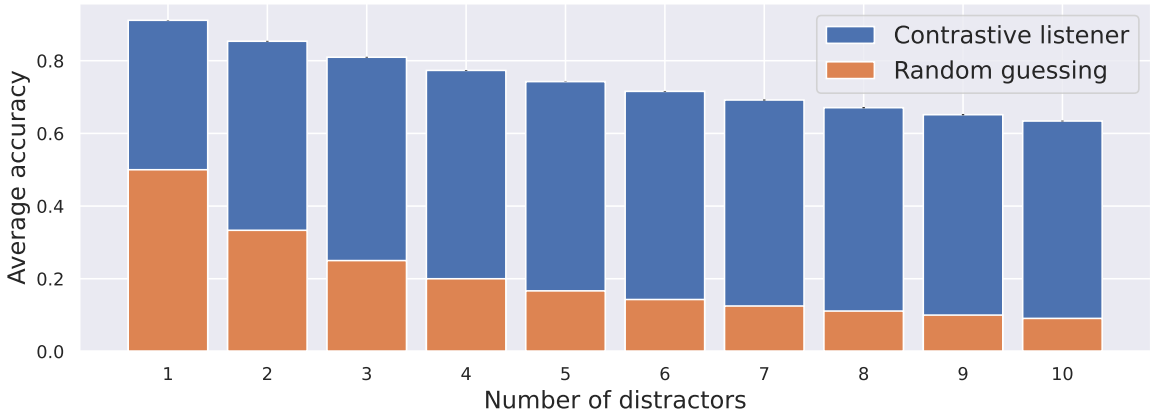

Figure 9. **Listening accuracy of a Transformer-based language encoder coupled with a ResNet-101 image-encoder, trained contrastively with Affection captions from scratch.**. The performance displayed is a function of the number of distractor images used at inference time and is the average resulting from five random seeds, used when pairing the target with randomly selected distractor images. Random guessing reflects performance when selecting the target uniformly at random. As expected, our neural listener fares significantly better, than random guessing, and also decreases its performance when more distractor images are considered.

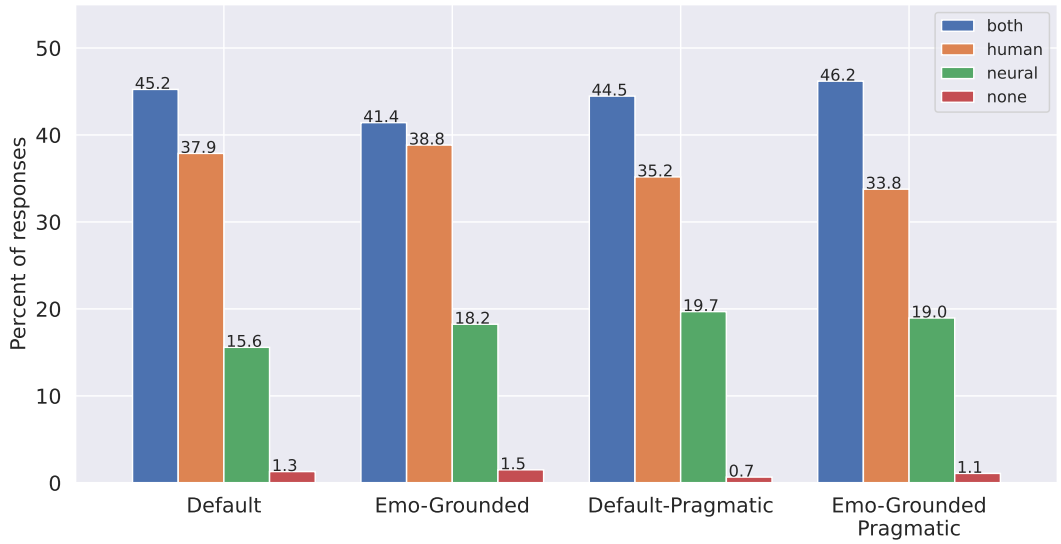

Figure 10. **Turing test results for our neural speaking variants.** For each variant we show the percent of its evaluated explanations that fall in one of the four categories ("both", "human", "neural" and "none") described in Section 6 of the main paper. All variants show strong baseline performance, with a minimum *aggregate* success rate ("both" and "neural" bars) of 59.6% attained by the emo-grounded variant, and a maximum rate of 65.2% attained by its (emo-grounded) pragmatic version. Note, that both pragmatic variants (two right-most histograms) outperformed their non-pragmatic versions.

Instructions

STEP 1: Look at the image and carefully read the two sentences.

then...
STEP 2: Decide which (if any) of the sentences could have been made **by a human**.

IMPORTANT.

1. The sentences are supposed to be **explanations** about **WHY** a human possibly felt, *or not*, any emotion upon seeing this image.
2. Sometimes humans *or* computers state explicitly the emotion they felt (e.g., happiness), but often, they do not!
3. Sometimes **BOTH** utterances will be from humans or computers.
4. **IGNORE** the **spelling** of the sentences and the lack of *punctuation*.
5. **IGNORE** **any** decision you made about **previously shown** image-sentence pairs in previously solved HITs **you did** for this task. I.e., **treat each HIT independently!**
6. **DO NOT submit more than ~50 HITS.**

Utterance A: this is a spectacular old building that i would love to explore

Utterance B: the architecture of the building is very intricate and detailed

Which utterance(s) were could be made by human(s), to justify a *possible* emotional reaction for the shown image?

- ○ Utterance **A**
- ○ Utterance **B**
- ○ **Both** could be made by humans
- ○ **None** is made by humans (both are made by a computer)

Figure 11. **User interface of emotional Turing test.** Upon reading the instructions (top) and observing the underlying image, each annotator had to select among the four options shown (bottom). In this example, the second utterance (B) is made by a neural speaker, while an annotator of Affection created the first utterance (A).

References

[1] Panos Achlioptas, Judy Fan, Robert XD Hawkins, Noah D Goodman, and Leonidas J. Guibas. ShapeGlot: Learning language for shape differentiation. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[2] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. ArtEmis: Affective language for visual art. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 2

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and Lawrence C. Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *Computing Research Repository (CoRR)*, abs/1504.00325, 2015. 1

[5] Z. Dave Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. *Computing Research Repository (CoRR)*, abs/1912.08830, 2019. 2

[6] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computing Research Repository (CoRR)*, abs/1512.03385, 2015. 1

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 4

[11] Ian Huang, , Panos Achlioptas, Tianyi Zhang, Sergey Tulyakov, Minhyuk Sung, and Guibas Leonidas. LADIS: Language disentanglement for 3D shape editing. In *Findings of Empirical Methods in Natural Language Processing*, 2022. 2

[12] 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. 7

[13] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[14] H. Kim, Y. Kim, S. J. Kim, and I. Lee. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 2018. 1, 2

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2017. 1

[16] S. Lo Piano. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Nature*, Humanities and Social Sciences Communications, 2020. 5

[17] A. Lopez. *Fdupes is a program for identifying or deleting duplicate files residing within specified directories.*, (accessed July 2022). Available at `https://github.com/adrianlopezroche/fdupes`. 1

[18] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[19] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Murphy Kevin. Generation and comprehension of unambiguous object descriptions. *Computing Research Repository (CoRR)*, abs/1511.02283, 2016. 1

[20] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. GRIT: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 4, 6

[21] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[22] You Quanzeng, Luo Jiebo, Jin Hailin, and Yang Jianchao. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *Computing Research Repository (CoRR)*, abs/1605.02677, 2016. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Computing Research Repository (CoRR)*, abs/2103.00020, 2021. 4

[24] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 1

[26] Amazon Mechanical Turk. *Acceptable Use Policy and Ethical Considerations*, (accessed November 2022). Available at `https://www.mturk.com/acceptable-use-policy`. 5

[27] Kees van Deemter. *Computational Models of Referring: A Study in Cognitive Science*. The MIT Press, 2016. 3

[28] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[29] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[30] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1989. 4

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 4

[32] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2014. 1

[33] Zanyar Zohourianshahzadi and Jugal K. Kalita. Neural attention for image captioning: Review of outstanding methods. *Computing Research Repository (CoRR)*, abs/2111.15015, 2021. 4