Emotional Theory of Mind: Assessing Vision and Language Models' Capabilities and Limitations

Author Names Omitted for Anonymous Review. Paper-ID 9

***Abstract*—How well can language models understand emotions in images? The Emotions in Context (EMOTIC) database is a challenging set of 23,571 images of people in various scenes and environments annotated with their apparent emotions. The emotional theory of mind problem is similar to a visual question and answering task, asking "How does the person in the bounding box feel?" Facial expressions, body pose, contextual understanding and implicit commonsense knowledge all contribute to the difficulty of the task, making emotion estimation in context currently one of the hardest problems in affective computing. The goal of this work is to evaluate the emotional knowledge embedded in recent vision language (CLIP) and large language models (GPT-3.5) on the EMOTIC dataset. In order to evaluate a purely text-based language model on images, we construct "narrative captions" relevant to emotion perception, using a set of 872 physical social signal descriptions related to 26 emotional categories, along with 224 labels for emotionally salient environmental contexts, sourced from writer's guides for character expressions and settings. We evaluate the use of the resulting captions in an image-to-language-to-emotion task. Experiments using zero-shot vision-language models on EMOTIC show that a gap remains in the emotional theory of mind task compared to prior work trained on the dataset, and that captioning with social signals and environment provides a better basis for emotion recognition than captions based only on activity. Limitations and opportunities for improvement are discussed.**

I. Introduction

Our ability to recognize emotions allow us to understand one another, build successful long-term social relationships, and interact in socially appropriate ways. Equipping our technologies with emotion recognition capabilities can help us improve and facilitate human-machine interactions [17]. However, emotion recognition systems today still suffer from poor performance [4] due to the complexity of the task. This innate and seemingly effortless capability requires understanding of the causal relations, contextual information, social relationships as well as using theory of mind for us to infer why someone might be feeling that way. Many image-based emotion recognition systems focus solely on using facial or body features [16, 26], which can have low accuracy in the absence of contextual information [3, 2].

The Affective Computing research community has been moving towards creating datasets and building models that includes or make use of contextual information. The EMOTIC dataset is a recent example that was introduced to address this problem, by including contextual and environmental factors to the recognition of emotions in still images [9]. It is found that the inclusion of these contextual information beyond facial features significantly improve the accuracy of the emotion recognition models [11, 14, 31]. However, making use of this information to infer the emotions of others requires commonsense knowledge and high-level cognitive capabilities such as reasoning and theory of mind [15].

Large Language Models (LLMs) that are based on the Transformer architecture [29] have been recently shown to excel at Natural Language Processing (NLP) tasks [6, 7] could provide us a way to achieve emotional theory of mind through linguistic descriptors. Providing an efficient way of processing sequenced data, and further introducing additional methods based on transformer encoder/decoder structures and pre-training techniques [8, 21], LLMs gained success in increasing accuracy and efficiency in NLP problems including multimodal tasks such as Visual Question Answering [1] and Caption Generation [30]. Recently, they have been also used in commonsense reasoning [24, 5, 12], emotional inference [13] and theory of mind [25] tasks, however their capabilities on emotional theory of mind in visual emotion recognition tasks have not been explored.

In this paper, we focus on the multi-label emotion recognition task from images with contextual information by creating a pipeline that includes caption generation and LLMs. We use vision language model (CLIP) to generate explainable linguistic descriptions of a target person in images using physical signals, action descriptors, gender and location information. We then use LLMs (GPT-3.5) to reason about the generated narrative text and predict the possible emotions that person might be feeling.

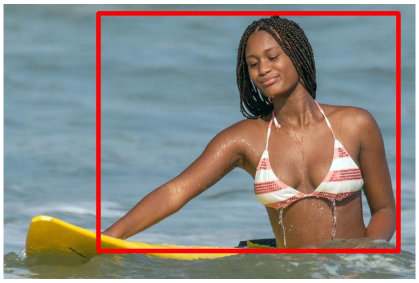
The contributions of this paper is as follows:

1) Investigate the emotion inference capabilities of vision language models and large language models (here, CLIP and GPT-3.5)
2) Introduce *narrative captions*, an interpretable textual representation for images of people experiencing an emotion
3) Evaluate more than 850 social cues to construct narrative captions and perform a large-scale visual emotional theory of mind task

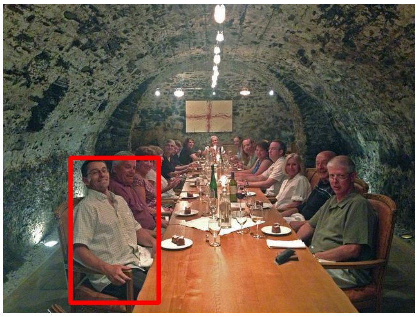
II. Methodology

In order to evaluate the capabilities of our pipeline, we provide performance comparisons with the zero-shot methods in the contextual emotion labeling task. We specifically investigate the vision language model CLIP [22], which can be used for out-of-the-box image classification, and GPT-3.5 for language reasoning. In order to access the natural language capabilities of GPT-3.5, we first aim to generate a caption of the image

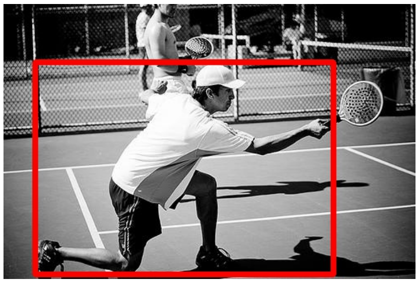
TABLE I
QUALITATIVE RESULTS



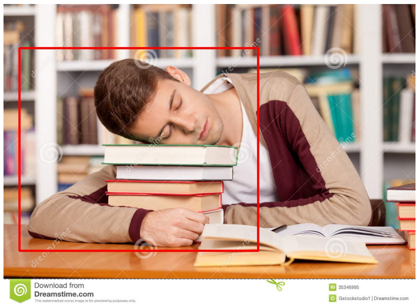
1. GT: happiness, peace, pleasure
NarraCap: This person is an older woman who is surfing water at a beach. She has saggy posture, bent neck. She has chin dipping to the chest, adopting a slumped posture. She is participating in relaxing activities.
GPT-3: confidence, engagement, happiness, peace, pleasure, sympathy
CLIP: happiness, sympathy, embarrassment, peace, confidence, esteem



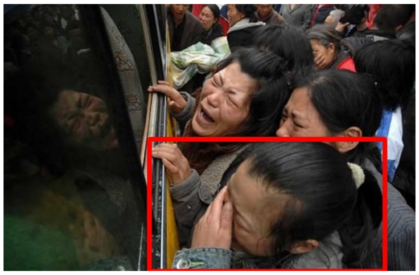
2. GT: anticipation, confidence, engagement, excitement, happiness, peace
NarraCap: This person is an older man who is tasting wine at a wine cellar. He has hands laying in the lap. He has fingers loosely clasped in one's lap. He is unbuttoning the top button of a shirt.
GPT-3: confidence, engagement, happiness, pleasure, esteem, anticipation
CLIP: disquietment, sadness, sympathy, doubt/confusion, surprise, peace



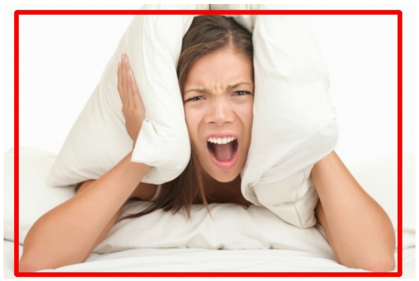
3. GT: anticipation, confidence, esteem
NarraCap: This person is an older man who is doing a tennis swing at a rec center. He has leg muscles tightening. He is clasping one's forearms, elbows pressed to the sides. He is using the body to nudge, push, or block (line jumpers).
GPT-3: confidence, engagement, happiness, pleasure, excitement, anticipation
CLIP: doubt/confusion, esteem, confidence, fatigue, sensitivity, suffering



4. GT: fatigue, peace
NarraCap: This person is an older man who is sleeping at a library. He is sleeping. He is curling up to take up less space. He is falling asleep in odd places at odd times.
GPT-3: fear, fatigue, disquietment, embarrassment, disconnection, sadness
CLIP: sympathy, engagement, suffering, disquietment, disconnection, fatigue



5. GT: disapproval, disquietment, fear, sadness, suffering
NarraCap: This person is an older woman who is huddling at a city bus. She is covering oneself (closing a jacket). She is squeezing the eyes shut, refusing to look. She is crying, wailing, begging for help.
GPT-3: fear, sadness, embarrassment, disquietment, anger, sympathy
CLIP: engagement, esteem, fear, sensitivity, sadness, sympathy



6. GT: affection, anger, annoyance, disapproval, disquietment, doubt/confusion, fear, sadness
NarraCap: This person is an older woman who is having a pillow fight at a run-down apartment. She is placing trembling fingertips against one's open mouth. She is squeezing the eyes shut, refusing to look. She is tossing and turning in bed, an inability to sleep.
GPT-3: Fear, annoyance, fatigue, embarrassment, sadness, sympathy
CLIP: aversion, sympathy, sensitivity, disquietment, anger, fatigue

such that the text representation allows for understanding the emotion of the person in it.

A. Narrative Captioning

We build upon previous captioning work that describes a person and their activity. First, given an image with the bounding box of a person, we extract the cropped bounding box and pass it along with the gender/age category (girl, boy, man, woman, older man, older woman) to CLIP to understand **who** is in the picture. Next, we pass the entire image to understand the **what** is happening in the image by analyzing the actions. The action list is extracted from the Kinetics-700 [27], UCF-101 [28], and HMDB datasets [10].

The contribution of the present *narrative caption* is to add the **how** aspect of the image. We obtained over 1000 social signals from a guide to writing about emotion [20]. We then filtered these signals to include only those visible in an image, resulting in 872 social signals. By passing these signals and the cropped bounding box to CLIP, we generate captions that describe the person in the image. To provide additional context, we use 224 environmental descriptors from a writer's guide to urban [19] and rural [18] settings to describe **where** the person in the scene is located.

B. Text to Emotion Inference

Using the caption, we then obtain a set of emotions experienced by this individual in the bounding box using a large language model, in this case GPT-3.5 (text-davinci-003). The prompt describes the emotions precisely as described in [9] and asks for the best six emotion labels understood from the caption :

<caption> *From suffering(which means psychological or emotional suffering; distressed; anguished), pain(which means physical pain), aversion(which means feeling disgust, dislike, repulsion; feeling hate), disapproval(which means feeling that something is wrong or reprehensible; contempt; hostile),*

[...] sadness(which means feeling unhappy, sorrow, disappointed, or discouraged), and sympathy(which means state of sharing others' emotions, goals or troubles; supportive; compassionate), pick a set of six most likely labels that this person is feeling at the same time. Six labels were requested from GPT-3.5 since the average number of ground truth labels in the validation set was 6.

III. Experiments

We use the Emotic dataset [9] which contains 23,571 images and 34,320 different people. Emotic covers 26 different labels as shown in Table II. The related emotion estimation task is to provide a list of emotion labels that matches those chosen by annotators. In approximately 25% of the person targets, the face was not visible, underscoring the role of context in estimating the emotion of a person in the image. Training set (70%) was annotated by 1 annotator, where Validation (10%) and Test (20%) sets were annotated by 5 and 3 annotators, respectively.

A. Evaluation Metrics and Baselines

Following the previous work in context-based emotion estimation, we use the standard Mean Average Precision(mAP) metric. We compare our method with the following methods.

Emotic Along with the Emotic dataset, Kosti et al. [9] introduced a two-branched network baseline. The first branch is a feature extraction module which gets the bounding box of the target person as input, and the second branch is an image feature extraction which gets the whole image as the input. The first branch is in charge of extracting body related features and the second branch is in charge of extracting scene-context features. Then a fusion network combines these features and estimates the output.

Emoticon Motivated by Frege's principle [23], Mittal et al. [14] proposed an approach by combining three different interpretations of context. They used pose and face features (context1), background information (context2), and interactions/social dynamics (context3). They used a depth map to model the social interactions in the images. Later they concatenated these different features and passed it to fusion model to generate outputs.

Random We consider selecting either 6 emotions randomly from all possible labels (**Rand**) or selecting 6 labels randomly where the weights are determined by the number of times each emotion is repeated in the validation set (**Rand(W)**).

Majority For the Majority baseline, we find the top 6 most common emotions in the validation set and use them as the predicted labels for all test images (**Maj**).

CLIP-only We evaluate the capabilities of the vision language model CLIP to predict the emotion labels. In this study we pass the image with the emotion labels in the format of: *"The person in the red bounding box is feeling {emotion label}"* to produce the probabilities that CLIP gives to each of these sentences. We then pick the 6 labels with the highest probabilities.

Action-only We consider only the **what** portion of the caption and disregard physical signals, environment and gender. We pass the set of actions to CLIP and generate a single action caption per image in the form of *This person is [activity]*. We then send this caption to GPT-3.5 and obtain 6 possible emotions using the same prompt template as described in Sec. 2B.

B. Results and Analyses

The results are shown in Table II, and example images and captions in Fig. I. It should be noted that Emotic and Emoticon utilized the EMOTIC training dataset to tune their models on the data, while the others do not directly utilize the EMOTIC training or validation set. We notice that anger prediction using NarraCap + GPT-3 was less effective than CLIP-only; one possible explanation is the abundance of acted/posed anger photos like the example in Table I-6. In anger photos (which may be rare and difficult to capture in the wild), pure emotion is being expressed without any particular activity, and therefore including a null action may improve the results. We further describe our ablations below.

How we should pick the physical signals? To answer this question we used different methods on the validation set. After getting probabilities from CLIP, We picked the top 1, 3, or 5 physical signals with highest probabilities to generate captions and record the mAP. As the second approach, we picked all the signals which had the probability greater than $Mean + std$, $Mean + 2 \times std$, $Mean + 5 \times std$, $Mean + 7 \times std$, and $Mean + 9 \times std$ as valid physical signals and used them to generate the captions. Note that the mean in this case is equal to 100 divided by number of physical signals. As you can see in Table III, the best results are for the when pick the top 3 physical signals. Note that we pass the cropped bounding box of the target to CLIP to get the physical signals.

How do different prompts affect the results? It is important to choose a good prompt for both GPT-3.5 and CLIP. The initial prompts we chose for CLIP was "*The person in the image is/has [gender/Action/physical signals]*" for gender, action, and physical signals, and "*The image is happening in a [location]*" for locations. The initial prompt for GPT-3.5 was the generated caption plus "*From suffering, pain, aversion, disapproval, anger, fear, annoyance, fatigue, disquietment, doubt/confusion, embarrassment, disconnection, affection, confidence, engagement, happiness, peace, pleasure, esteem, excitement, anticipation, yearning, sensitivity, surprise, sadness, and sympathy, pick top labels that describe the emotion of this person.*" We randomly pick 100 samples of validation data and run the experiment which these prompt on it. We obtained a mAP equal to 26.03.

We then tried to change the GPT-3.5 prompt to add the meaning of labels provided by EMOTIC (see Section 2.B) and we obtained a mAP equal to 26.27. The other prompt we used

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON EMOTIC DATASET. ACTION AND NARRACAP CAPTIONING USE GPT-3 FOR INFERENCE.

Emotions	Emotic	Emoticon	Rand	Maj	Rand(W)	CLIP	Action	NarraCap
Affection	27.85	45.23	13.27	13.41	13.58	22.63	18.13	19.11
Anger	09.49	15.46	2.45	2.45	2.42	6.09	2.85	3.63
Annoyance	14.06	21.92	4.74	4.89	4.91	5.14	5.75	6.26
Anticipation	58.64	72.12	46.86	47.05	47.23	47.31	48.91	48.76
Aversion	07.48	17.81	3.05	3.05	3.07	3.03	3.38	3.22
Confidence	78.35	68.65	47.33	47.43	47.84	52.12	51.78	49.68
Disapproval	14.97	19.82	4.61	4.55	4.54	4.44	4.77	4.62
Disconnection	21.32	43.12	15.80	15.54	15.34	15.42	15.82	17.27
Disquietment	16.89	18.73	12.45	12.43	12.25	13.40	12.70	13.20
Doubt/Confusion	29.63	35.12	14.31	14.52	14.60	14.58	14.60	14.92
Embarrassment	03.18	14.37	1.89	1.88	1.85	1.96	2.07	2.19
Engagement	87.53	91.12	77.28	77.33	77.24	76.57	79.02	77.85
Esteem	17.73	23.62	13.33	13.64	14.03	13.75	13.81	14.17
Excitement	77.16	83.26	47.80	47.94	47.34	48.31	53.31	54.41
Fatigue	09.70	16.23	6.02	5.91	5.95	7.38	6.36	6.52
Fear	14.14	23.65	3.27	3.20	3.20	3.67	3.56	4.08
Happiness	58.26	74.71	48.05	47.95	48.27	52.60	48.93	52.52
Pain	08.94	13.21	2.03	1.96	1.95	1.99	2.11	2.86
Peace	21.56	34.27	14.70	14.61	14.33	16.38	16.15	16.79
Pleasure	45.46	65.53	29.17	28.99	29.35	30.74	29.50	31.41
Sadness	19.66	23.41	5.32	5.34	5.40	8.20	6.22	8.63
Sensitivity	09.28	8.32	2.88	2.93	2.94	3.42	2.95	2.92
Suffering	18.84	26.39	3.50	3.50	3.47	3.06	3.50	3.50
Surprise	18.81	17.37	6.34	6.14	6.15	6.87	6.11	6.11
Sympathy	14.71	34.28	8.33	8.45	8.34	9.01	8.52	9.55
Yearning	08.34	14.29	6.48	6.47	6.45	6.40	6.58	6.50
mAP	27.38	35.48	16.97	16.98	17.00	18.25	17.98	18.49

TABLE III
RESULTS FOR DIFFERENT APPROACHES OF PICKING PHYSICAL SIGNALS ON THE 100 RANDOMLY PICKED PEOPLE FROM VALIDATION SET. WE ALSO REPORT THE AVERAGE NUMBER OF PHYSICAL SIGNALS AND AVERAGE NUMBER OF WORDS IN CAPTIONS FOR EACH APPROACH.

	mAP	ave #sigs	ave cap len
top1	25.99	1.0	18.83
top3	**26.27**	**3.0**	**35.52**
top5	25.36	5.0	52.06
$mean + std$	25.21	37.99	315.2
$mean + 2 \times std$	25.39	18.29	162.18
$mean + 5 \times std$	25.84	4.98	52.87
$mean + 7 \times std$	26.184	3.0	35.84
$mean + 9 \times std$	25.96	1.82	25.62

was to ask *pick top **six** labels that describe the emotion of this person* which resulted in a mAP of 26.20. After changing the prompt to *pick a set of six most likely labels that this person is feeling at the same time*, mAP increased to 26.63.

IV. CONCLUSION

In this paper, we begin to explore the capabilities of vision language models and LLMs (here, CLIP and GPT-3.5) for the visual emotional theory of mind task. We first find that our results were inferior to Emotic and Emoticon baseline models which were trained specifically for this task. Nevertheless, we observe the potential of the LLM, as it outperforms the self-supervised vision language model CLIP on its own. Further research could explore improving the narrative caption by adding other contextual factors to the caption, such as human-object interactions and relationships with others. Other improvements could involve more carefully selecting physical signals by body part, depending on their prominence in the scene. Further studies using alternative captioning methods and GPT-3.5 should be done. Multimodal reasoning models such as GPT-4 may also perform better by jointly performing the reasoning and vision processing tasks.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings*

of the IEEE international conference on computer vision, pages 2425–2433, 2015.

[2] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.

[3] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current directions in psychological science*, 20(5):286–290, 2011.

[4] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[5] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[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] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019.

[10] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[11] Nhat Le, Khanh Nguyen, Anh Nguyen, and Bac Le. Global-local attention for emotion recognition. *Neural Computing and Applications*, 34(24):21625–21639, 2022.

[12] Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, 2022.

[13] Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 2022.

[14] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[15] Desmond C Ong, Jamil Zaki, and Noah D Goodman. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357, 2019.

[16] Maja Pantic and Leon JM Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[17] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[18] Becca Puglisi and Angela Ackerman. *The Rural Setting Thesaurus: A Writer's Guide to Personal and Natural Places*, volume 4. JADD Publishing, 2016.

[19] Becca Puglisi and Angela Ackerman. *The Urban Setting Thesaurus: A Writer's Guide to City Spaces*, volume 5. JADD Publishing, 2016.

[20] Becca Puglisi and Angela Ackerman. *The emotion thesaurus: A writer's guide to character expression*, volume 1. JADD Publishing, 2019.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Michael David Resnik. The context principle in frege's philosophy. *Philosophy and Phenomenological Research*, 27(3):356–365, 1967.

[24] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[25] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

[26] Konrad Schindler, Luc Van Gool, and Beatrice De Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9): 1238–1246, 2008.

[27] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

[28] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob

Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[31] Zili Wang, Lingjie Lao, Xiaoya Zhang, Yong Li, Tong Zhang, and Zhen Cui. Context-dependent emotion recognition. *Journal of Visual Communication and Image Representation*, 89:103679, 2022.