Decoding Emotions in Abstract Art: Cognitive Plausibility of CLIP in Recognizing Color-Emotion Associations

Hanna-Sophia Widhoelzl (hannasophia.widhoelzl@gmail.com)
Institute for Interdisciplinary Studies, University of Amsterdam
Amsterdam, the Netherlands

Ece Takmaz (e.k.takmaz@uu.nl)
Department of Information and Computing Sciences, Utrecht University
Utrecht, the Netherlands

Abstract

This study investigates the cognitive plausibility of a pretrained multimodal model, CLIP, in recognizing emotions evoked by abstract visual art. We employ a dataset comprising images with associated emotion labels and textual rationales of these labels provided by human annotators. We perform linguistic analyses of rationales, zero-shot emotion classification of images and rationales, apply similarity-based prediction of emotion, and investigate color-emotion associations. The relatively low, yet above baseline, accuracy in recognizing emotion for abstract images and rationales suggests that CLIP decodes emotional complexities in a manner not well aligned with human cognitive processes. Furthermore, we explore color-emotion interactions in images and rationales. Expected color-emotion associations, such as red relating to anger, are identified in images and texts annotated with emotion labels by both humans and CLIP, with the latter showing even stronger interactions. Our results highlight the disparity between human processing and machine processing when connecting image features and emotions.

Keywords: emotion perception; abstract art; color-emotion interaction; cognitive plausibility; CLIP

Introduction

Abstract art's subjective and open-ended nature evokes intricate emotional responses in viewers. Its intrinsic ambiguity in both perception and interpretation has intrigued artists and researchers alike: abstract art presents an idiosyncratic challenge to our understanding of visual communication. Understanding emotions in abstract art is a complex aspect of human cognition, affected by individuals' education, cultural background, and social context. Previous research showed that features such as strokes, shapes, and visual harmony influence someone's sentiment (Ko, Yoon, & Kim, 2016; Lu et al., 2012; Sartori, Yanulevskaya, et al., 2015; Zhao et al., 2014). Most studies focused on the interaction between color and emotion. While color-emotion associations vary with cultural factors and subjective interpretations, psychological studies identified universal, yet complex trends: Brighter colors such as yellow are associated with happiness, while red may signify anger. Blue is commonly linked to sadness, black to fear, and grey to boredom and depressive moods. Brown and green are often thought to evoke disgust (Hemphill, 1996; Liu, 2022; Sutton & Altarriba, 2016).

A novel frontier in this domain is the intersection of artificial intelligence (AI) and human cognition. Recent advancements in Natural Language Processing (NLP) and Computer Vision (CV) gave rise to pretrained models such as CLIP (Contrastive Language-Image Pretraining) developed by OpenAI (Radford et al., 2021). CLIP excels at language understanding and zero-shot image classification across diverse content types including realistic photographs, cartoons, and diagrams.[1] CLIP was later integrated into various models that leverage its cross-modal understanding for downstream tasks (Agarwal et al., 2021), such as generating images from textual prompts (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022), describing images with natural language (Li, Li, Savarese, & Hoi, 2023; Berrios, Mittal, Thrush, Kiela, & Singh, 2023), answering questions related to images (Shen et al., 2022), and emotion recognition in naturalistic images (Bondielli, Passaro, et al., 2021). The cognitive plausibility of such models, and the degree to which their behavior mirrors human cognitive processes, is a growing area of interest. The assumption is that these models capture key elements of human-like understanding, aiding in explaining aspects of human cognition and behavior (Kennedy, 2009; Beinborn & Hollenstein, 2024). Uncovering potential disparities and shortcomings in the models' performance in emotion recognition in abstract art compared to human expectations contributes to creating more accurate artificial models of human cognitive processes. For instance, Phillips and Pearl (2015) argued that incorporating more cognitive plausible assumptions into computational models of language acquisition improved the model's performance and explanatory power.

This research leverages CLIP to explore the potential of current state-of-the-art vision and language models in elucidating the intricate cognitive processes by which abstract art elicits emotions. We investigate whether the representations generated by CLIP can be effective in predicting these emotional responses, assessing cognitive plausibility by evaluating model outputs. We employ the FeelingBlue dataset, a multimodal corpus designed to scrutinize the emotional implications of color in language and art (Ananthram, Winn, & Muresan, 2023). The dataset encompasses abstract art images and textual justifications for the emotions assigned to those images. We investigate factors affecting CLIP's emotion recognition abilities, such as model size and types of prompts, concentrating on color-emotion associations.

Our hypotheses are: (1) CLIP will demonstrate accuracy

[1] In a zero-shot classification setup, a model identifies and categorizes content not encountered during training. This tests the model's capacity to generalize its understanding to novel data.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

above chance level (five emotions, 20%) in predicting emotions for abstract art images and textual justifications. (2) Using a similarity-based encoding approach with CLIP embeddings to approximate the emotion labels of novel stimuli, we expect to see improved emotion recognition in images, providing insights into the model's nuanced understanding of emotions. (3) Analyzing color-emotion interactions in images and textual justifications provided by annotators rationalizing their emotion allocation is expected to corroborate existing literature on color-emotion associations.

Related Work

Sentiment analysis of textual content has received extensive attention in the realm of NLP (Dang, Moreno-García, & De la Prieta, 2020; Nasukawa & Yi, 2003; Solangi et al., 2018). However, the exploration of affective properties in images is relatively limited. Traditional CV generally focuses on predicting emotions from images of facial expressions, body postures, and gestures. In contrast, this study aims to understand emotions evoked by conceptual scenes, making it a distinct task. While both textual and visual content can be inherently implicit and ambiguous, the abstract nature of utilized images may render emotion classification more complex. Nevertheless, studies reported that images can potentially elicit stronger emotional impacts on humans than text (Casas & Williams, 2019). Consequently, the significance of training multimodal models to reach a better representation of human behavior and cognition is increasingly recognized.

Considering CLIP, a notable study by Bondielli et al. (2021) delved into the model's behavior in emotion classification. CLIP exhibited promising results for recognizing emotions in naturalistic images under zero-shot settings. Fine-tuning CLIP improved performance substantially, emphasizing CLIP's capacity to effectively learn from visual and textual cues. This research contributes to understanding CLIP's adaptability to emotion-related tasks and its ability to capture affective aspects of visual content.

There is limited research focusing on AI models' cognitive plausibility for emotion recognition in abstract art. Yanulevskaya et al. (2008) confirmed the potential of Support Vector Machines in perceiving emotions elicited through art by using a scene categorization system, while emphasizing the impact of painting techniques and color on emotion distinction. Their dataset depicted objects, people, and landscapes rather than consisting exclusively of fully abstract pieces. Thus, our study builds upon existing literature to explore the challenges and potential advancements of models in mirroring human emotion perception in abstract images.

Dataset

The FeelingBlue corpus (Ananthram et al., 2023) explores the implications of changes in the colors of abstract paintings on emotional connotations in various contexts, mediated by visual elements such as lines, strokes, shapes, and language. It comprises images sourced from WikiArt (S. Mohammad & Kiritchenko, 2018) and DeviantArt (Sartori, Culibrk, Yan, & Sebe, 2015). Identifiable objects (e.g., flowers, statues) were excluded to eliminate confounding factors on perceived emotions. The corpus contains five emotion subsets: anger, disgust, fear, happiness, and sadness. FeelingBlue consists of 19,788 randomly generated 4-tuples of abstract art images. Annotators systematically ranked the images within the tuples based on a given emotion label according to Best-Worst Scaling (BWS) (Flynn & Marley, 2014). BWS involves annotators selecting the 'best' and 'worst' options from a set, revealing relative preferences across the dataset. In FeelingBlue, annotators were instructed to select images that most and least evoked the stated emotion within a tuple. This might lead to overlapping emotion subsets, where one image may be labeled with different emotions when presented in different contexts (Figure 1). Subsequently, annotators provided textual justifications, referred to as 'rationales', for their choices of the 'least' and 'most' ranked image for evoking the specified emotion by describing salient visual features.

Methods

Model

CLIP (Radford et al., 2021) was trained on 400 million image-text pairs derived from the web. Its architecture involves an image encoder and a text encoder to map images and text into a shared vector space. During pretraining, CLIP's contrastive training objective aims to increase the alignment score between a given image from the dataset and the text associated with this image (its caption). CLIP's ability to associate various visual concepts with corresponding captions generalizes to robust zero-shot performance in a plethora of vision and language tasks, including image and text emotion classification. Thus, CLIP seems a potentially valuable tool for exploring emotion recognition in abstract images. We employ two pretrained versions of CLIP's visual encoder: ViT-B/32 and ViT-L/14, based on the Vision Transformer architecture (Dosovitskiy et al., 2021). ViT-L/14, while significantly larger and demanding greater computational resources, offers potentially superior performance to B/32. To assess CLIP's performance in subsequent analyses, we compute a score that measures the scaled cosine similarity between the representation of the input (images or rationales; i) and emotion labels as captions (c) in the form of vectors, which is called CLIPScore ($\text{CLIPScore}(i,c) = 2.5 * \max(\cos(i,c),0)$) (Hessel, Holtzman, Forbes, Le Bras, & Choi, 2021). CLIPScore has been proposed as a metric to evaluate model-generated image captions, outputting scores well-aligned with human judgments.

Analyses

Data Preprocessing We preprocess the FeelingBlue corpus to generate an image-emotion dictionary that facilitates further analyses and interpretations of results. In the original dataset, individual images are evaluated by multiple annotators on multiple emotions. For instance, image_328 is

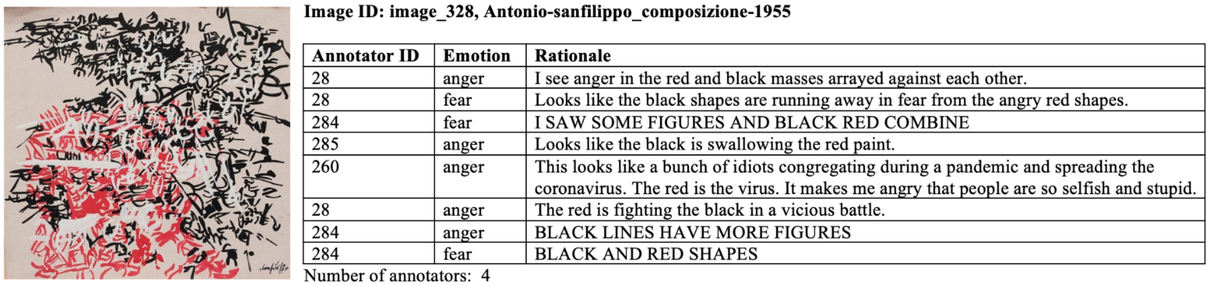

Image ID: image_328, Antonio-sanfilippo_composizione-1955

Annotator ID	Emotion	Rationale
28	anger	I see anger in the red and black masses arrayed against each other.
28	fear	Looks like the black shapes are running away in fear from the angry red shapes.
284	fear	I SAW SOME FIGURES AND BLACK RED COMBINE
285	anger	Looks like the black is swallowing the red paint.
260	anger	This looks like a bunch of idiots congregating during a pandemic and spreading the coronavirus. The red is the virus. It makes me angry that people are so selfish and stupid.
28	anger	The red is fighting the black in a vicious battle.
284	anger	BLACK LINES HAVE MORE FIGURES
284	fear	BLACK AND RED SHAPES

Number of annotators: 4

Figure 1: Emotion allocation for 'Composizione' by Antonio Sanfilippo (1955), from FeelingBlue (Ananthram et al., 2023).

ranked as evoking the most 'anger' five times and the most 'fear' three times within different 4-tuple image groups. Four different annotators assessed image_328 (Figure 1). Participants could encounter the same images multiple times in distinct image groups or under different emotion label queries. We assign the emotion mentioned most frequently, in this case, 'anger', to each image to create the aforementioned dictionary featuring 901 image-emotion pairs.[2] While the corpus initially contains 934 images, 33 images could not be assigned an emotion as they do not rank highest within any given tuple, prompting their removal from our dataset. Among the remaining images, 30.3% are labeled as 'happiness', 9.5% as 'anger', 8.8% as 'sadness', 16.0% as 'fear', and 35.4% as 'disgust'. The rationales are preprocessed using NLTK (Loper & Bird, 2002). All text is converted to lowercase. British spelling is converted to American spelling.

Linguistic Features of Rationales We analyze the most common adjectives describing images linked to specific emotions mentioned in the rationales, focusing on color-related terms (9,912 rationales in total). Rationales for images rated as evoking anger, sadness, happiness, fear, or disgust are accordingly grouped by these emotions.

Zero-Shot Emotion Classification We test CLIP's zero-shot ability in two setups using ViT-B/32 and ViT-L/14: (1) rationale to emotion label, (2) image to emotion label. In the former, rationales serve as textual input, and CLIP is presented with various emotion prompts. Thus, we compare text to text here by considering their similarity in CLIP's textual representation space, leveraging the model's robust understanding of textual semantics. In the latter, we compare each image to the emotion prompts. Thus, we compare text to image here, mirroring CLIP's contrastive learning objective.

As reported by Bondielli et al. (2021), CLIP is sensitive to the wording of its prompts describing the emotion labels. We reuse their best-performing prompt structure, referred to as 'standard prompts', which puts the emotion label in context. For control purposes, we include 'single-word prompts'. Finally, we create 'art-tailored prompts' to accommodate the nature of the stimuli by tailoring prompts to align with the specific themes of abstract art. We designed them through careful consideration of the distinctive features inherent in abstract art by stating the emotion label indirectly through artistic expressions. Terms representing the emotion labels are chosen based on the NRC Word-Emotion Association Lexicon (EmoLex) (S. M. Mohammad & Turney, 2013). EmoLex offers a list of English words associated with basic emotions identified through crowd sourcing, such as 'melancholy' linked to 'sadness'. This results in three sets of prompt structures with five different prompts, each representing one emotion. Firstly, the 'Standard' prompt: {a sentence|an image} that elicits {anger|happiness|sadness|fear|disgust}. Secondly, the 'Single-Word' prompt: {anger|happiness|sadness|fear|disgust}. Lastly, the 'Art-Tailored' prompt: {a sentence|an image} that evokes a sense of fiery intensity *(anger)*; {a sentence|an image} that captures a burst of vibrant joy *(happiness)*; {a sentence|an image} that conveys a profound sense of melancholy *(sadness)*; {a sentence|an image} that instils a feeling of eerie unease *(fear)*; {a sentence|an image} that evokes a sense of unsettling chaos *(disgust)*.

The prompt resulting in the highest CLIPScore when matched with a rationale or image is assumed to have the best fit with the input and is thus used to assign an emotion label. For example, if the prompt representing 'happiness' achieves the highest CLIPScore when matching the rationale 'the sun is bright' with a set of prompts, the rationale is predicted to correspond to 'happiness'. These predicted labels are subsequently compared to the emotion labels assigned by human annotators to calculate CLIP's accuracy.

Similarity-Based Emotion Prediction We adopt a similarity-based encoding approach to further investigate the extent to which CLIP mirrors human emotion perception tendencies. The approach is inspired by Anderson, Zinszer, and Raizada (2016), who reported comparable accuracy to regression-based methods while obviating the need for model fitting when predicting fMRI signals for novel stimuli. The authors calculate the weighted average of the signals induced by a neighboring set of stimuli in the representational space to approximate the signals for a novel stimulus (Anderson et al., 2016). In our work, we expect that images evoking

[2] We picked the most frequently mentioned emotion as a simplification to facilitate further analyses; however, we acknowledge the complexity of emotional responses to images, which may simultaneously evoke various emotions or differ among viewers.

similar emotions have similar representations in the shared feature space. Hence, we assess if images with similar representations by CLIP help predict the emotion label of the target stimulus. We divide the dataset into 80% training set and 20% test set. Emotion labels in the training set are transformed into one-hot vectors. We utilize the ten most similar training images for each test image by using cosine similarity between CLIP-encoded images. Using the ten most similar images is a trade-off between having a rich pool of samples to base the predictions on and computational efficiency. Subsequently, we calculate the weighted averages of the ten most similar images' emotion labels for each test image to predict its emotion label.

Color-Emotion Interaction Since literature suggests an apparent interaction between emotions and colors, we assess how color words mentioned in the rationales relate to emotions by probing for the following words that we identified through the linguistic features analysis: 'black', 'dark', 'red', 'bright', 'yellow', 'white', 'blue', 'pink', 'orange', 'colorful, 'grey', 'green', 'brown', and 'purple'. Firstly, we plot these words based on the emotions assigned to rationales by human annotators. We then examine the distribution of color terms in rationales based on CLIP's predicted emotion labels. Due to computational constraints, we focused on ViT-B/32. We also analyze the interaction between emotions and the colors in images. Color distributions per human-annotated emotion are compared to CLIP ViT-B/32's emotion classification. To facilitate our analyses, we identify the most dominant color of images using ColorThief.[3] We map the resulting RGB values to color terms through Webcolors[4] (CSS Color Module Level 3). These color terms, such as 'lime', were then grouped into broader categories, such as 'green'.

Results & Discussion

Images

Zero-Shot Emotion Classification CLIP's accuracy for emotion recognition in images only slightly surpasses chance level under zero-shot settings (Table 1), which is lower than anticipated, yet still supporting hypothesis (1). The highest accuracy is 29.12% with ViT-L/14 and 'art-tailored prompts'. The lowest accuracy of 11.56% is observed with ViT-B/32 and 'single-word' prompts. CLIP's performance is fairly sensitive to prompt wording: Providing the to-be-predicted emotion label in context leads to superior performance over stating the emotion in isolation (Bondielli et al., 2021). Utilizing prompts tailored to art achieves the highest performance, suggesting that the context in prompts should relate to the nature of the images at hand. Yet ideally, models should attain a level of proficiency where they can be considered models of human cognition or used as tools for investigating human cognition, even when provided with single-word prompts. ViT-L/14 consistently outperforms B/32 as expected, given its larger model size and enhanced generalization abilities.

CLIP demonstrated lower accuracy in emotion recognition for abstract art than for naturalistic images (Bondielli et al., 2021). Notably, the dataset used in the referenced study included images featuring individuals expressing emotions through facial expressions or body language. These additional cues might have contributed to higher accuracy. Additionally, differences in considered emotion classes and evaluation metrics require careful interpretation of the results when drawing direct comparisons with our findings. Nevertheless, this suggests that CLIP encodes emotional concepts in abstract images differently from naturalistic images.

Table 1: Emotion recognition accuracy (%) across data types, model versions, and prompt sets.

Data Type	CLIP Model	Prompts	Accuracy
Images	ViT-B/32	Single-Word	11.35
		Standard	11.56
		Art-Tailored	**24.73**
	ViT-L/14	Single-Word	14.12
		Standard	20.24
		Art-Tailored	**29.12**
Rationales	ViT-B/32	Single-Word	34.79
		Standard	**41.99**
		Art-Tailored	26.85
	ViT-L/14	Single-Word	**45.04**
		Standard	44.71
		Art-Tailored	27.41

Similarity-Based Emotion Prediction Since CLIP ViT-L/14 did not achieve considerably higher accuracy than B/32 in the zero-shot experiments, we exclusively applied similarity-encoding to B/32 embeddings to predict images' emotion labels. This resulted in an accuracy of 47.51% on the test set, demonstrating superiority over random guessing and zero-shot performance, thereby corroborating hypothesis (2). This underscores the capability of CLIP's embeddings to decipher meaningful features associated with emotional content, as CLIP arrived at more correct predictions when considering similarly encoded images. Utilizing multiple similar samples enhances contextual awareness and potentially enables capturing more nuanced emotional subtleties that might be missed during zero-shot classification. This shows the importance of contextual awareness in emotion recognition of abstract art, highlighting the inherent strength of human cognition in a task that proves more challenging for machines when relying solely on visual cues, without access to the sociocultural environment (Mesquita & Boiger, 2014; Mittal, Bera, & Manocha, 2021; Yang et al., 2022).

Despite this improved accuracy, our findings suggest that CLIP's encoding of abstract images is not highly indicative of their emotional status. Alternatively, features of images eliciting the same emotion might vary considerably, thereby impeding CLIP's emotion classification ability. Another pos-

[3] https://lokeshdhakar.com/projects/color-thief/

[4] https://webcolors.readthedocs.io/en/latest/

sible, yet contradicting, hindrance is that features of images evoking different emotions might overlap, given the lack of concrete emotional cues.

Color-Emotion Interaction We explore the distribution of dominant colors of images per emotion label to identify color-emotion associations. Firstly, considering each image's dominant color and human-annotated emotion (Figure 2), we observe patterns aligning partially with our hypothesis (3). The dominant colors of 'happiness' images are often orange, white, and yellow. Images of negative emotions, especially 'fear', include black. The 'anger' class has the highest proportion of red images, though not considerably more than 'happiness'. Surprisingly, 'happiness' images are more often blue than 'sadness' images. Similarly, although 'disgust' images feature green or brown, we also see a large proportion of grey, as in all emotion categories. Instead, 'fear' images have the highest proportion of green. 'Happiness' and 'disgust' images show comparable amounts of brown.

Different patterns of color-emotion distributions emerge in CLIP's emotion classification for all included images, meaning both correctly and incorrectly labeled images, based on 'art-tailored' prompts (Figure 3). Firstly, 'happiness' images are frequently white. 'Fear', 'disgust', and 'anger' images have a considerable number of grey and black color. 'Anger' images have the highest proportion of red. While 'happiness' images feature the highest proportion of green, the 'disgust' category has the most green images among the negative emotions. Contrary to our hypotheses, the largest proportion of blue images is classified as 'fear' instead of 'sadness'. 'Anger' images have the largest proportion of brown instead of 'disgust' images. Notably, CLIP categorizes only one image as 'sadness'. The 'sadness'-color distribution is therefore unrepresentative and should be disregarded.

Interestingly, CLIP's classification of images to emotions based on dominant colors aligns more closely with the theories from literature than the associations found between colors and human-annotated emotions. That CLIP's emotion recognition has not achieved high accuracy, however, indicates that factors such as shapes, lines, textures, or overall compositions influenced the annotators' choices for emotion labeling. For CLIP, these factors might have proven more difficult to integrate into a holistic understanding of the images' emotions.

Rationales

Linguistic Features of Rationales Linguistic analyses of rationales reveal associations between emotions and descriptive language, including color-emotion interactions. Among images associated with happiness, commonly used adjectives are 'attractive', 'bright', and 'colorful'. These adjectives show limited overlap with those frequently found in annotations describing images associated with negative emotions, namely 'red', 'black', 'dark', and 'rough'.

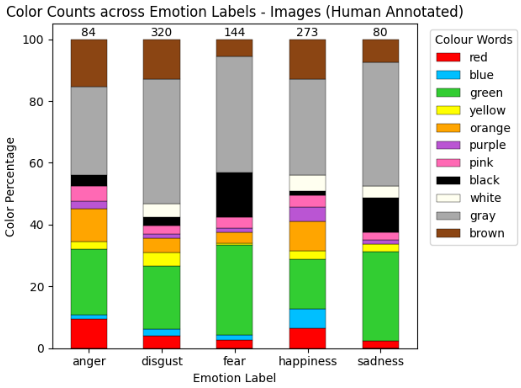

Figure 2: Color-emotion associations in human-annotated images with frequency counts per emotion category.

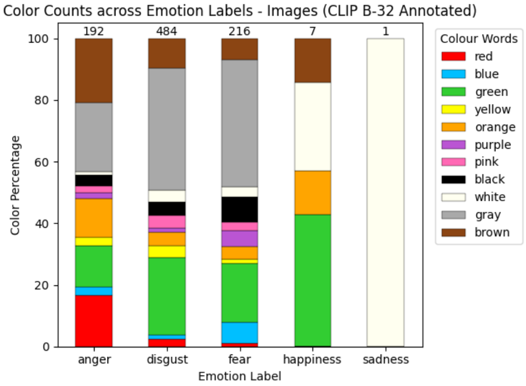

Figure 3: Color-emotion associations in CLIP-annotated images with frequency counts per emotion category.

Zero-Shot Emotion Classification CLIP's zero-shot emotion recognition performance in rationales is considerably higher than in images with an accuracy of 45.04% with ViT-L/14 and 'single-word' prompts. Several factors might contribute to this. Firstly, textual justifications offer more explicit emotional cues, enabling CLIP to discern emotions more effectively compared to abstract images. CLIP may have learned linguistic patterns associated with certain emotions. Again, ViT-L/14 outperforms B/32 for all prompts. The lowest performance is attained with 'art-tailored' prompts, revealing an opposite pattern as found with images. This might be due to rationales providing context through concrete and abstract words. Thus, a concise prompt might be easier for CLIP to match with rationales. Abstract art images, instead, lack concrete concepts. CLIP might thus benefit from additional context, while humans might naturally derive emotionally meaningful interpretations from abstract art.

Color-Emotion Interaction Examining the distribution of color words mentioned in rationales based on human annotations reveals the following patterns (Figure 4): Words such as ‘bright’ and ‘colorful’ are commonly found in rationales describing ‘happy’ images. ‘Black’ and ‘dark’ are prevalent in rationales labeled with negative emotions. ‘Blue’ is more frequent in ‘sadness’ rationales than other negative emotions, while ‘red’ is associated with ‘anger’. ‘Green’ and ‘brown’ appear more often in ‘disgust’ rationales, yet these color-emotion associations are less pronounced. A substantial proportion of ‘fear’ rationales contain ‘black’, but the proportion of ‘black’ in ‘sadness’ rationales is even larger. These findings align with our initial hypotheses, yet highlight the complexity of color-emotion interactions.

Results of the color word distributions in rationales according to CLIP’s classification can be seen in Figure 5. Again, ‘happiness’ rationales contain higher proportions of words such as ‘bright’ or ‘colorful’. Negative emotions are associated with ‘black’ and ‘dark’. The associations between ‘anger’ and ‘red’, ‘sadness’ and ‘blue’, and ‘disgust’ and ‘green’ and ‘brown’ are more pronounced than in the human-annotated pattern. Overall, these patterns mostly support our hypothesis (3), which reflects the relatively high accuracy of rationale-emotion labeling with CLIP.

CLIP’s potential to recognize color-emotion associations in both images and text validates the relevance of color in emotional connotations. As found with images, CLIP’s categorization of rationales to emotions adheres to established color-emotion associations more strongly than human annotations. Hence, even for rationales mentioning explicit color words, other factors affect emotion perception in humans.

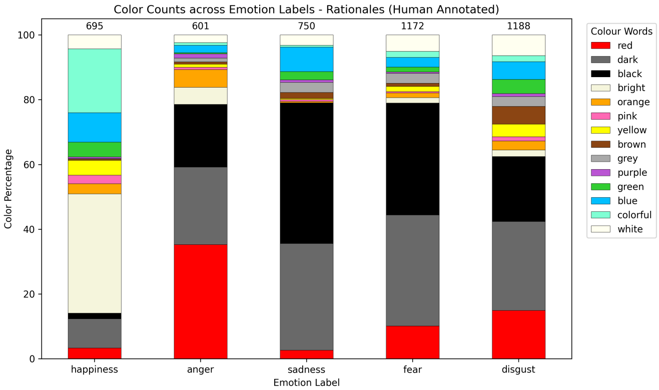

Figure 4: Color-emotion associations in human-annotated rationales with frequency counts per emotion category.

General Discussion

This study contributes to the growing body of research at the intersection of AI and art, paving the way for future exploration of human cognition and emotion through the lens of AI models. Exploring CLIP as a proxy model unveiled promising prospects for leveraging advanced vision and language models in decoding the emotional complexities of abstract art and textual justifications while reflecting on distinctive strengths of human cognition, such as context awareness.

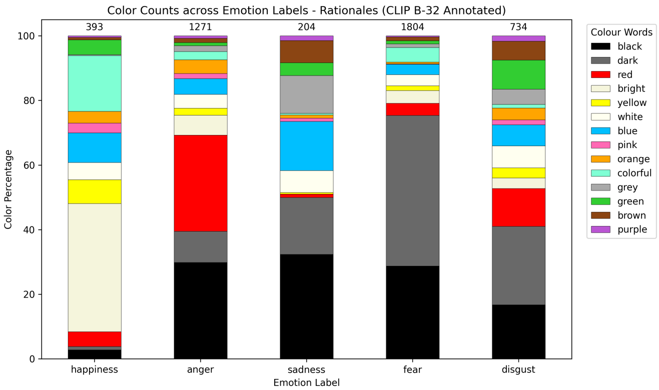

Figure 5: Color-emotion associations in CLIP-annotated rationales with frequency counts per emotion category.

Overall, CLIP seems to exhibit difficulties bridging the affective gap between low-level features and abstract emotions, as there was no convincing evidence that CLIP latently encodes emotion elicited by images rigorously. This suggests potential limitations in its cognitive plausibility for tasks related to understanding and representing complex emotional states. That CLIP recognizes emotions in abstract art less accurately than in naturalistic images under zero-shot settings may be attributed to the inherent challenges of abstract art’s subjective and ambiguous nature. Abstract art often lacks recognizable objects or scenes that may provide apparent emotional context, which may have hindered CLIP’s ability to identify subtle emotional nuances and fine-grained details. Humans excel at capturing nuanced emotional expressions, leveraging a broader contextual understanding inherent in human cognition, such as personal experience, cultural influences, or recognition of familiar objects and scenes (Barrett, Mesquita, & Gendron, 2011; Hess & Hareli, 2016). These challenges faced by CLIP highlight a potential disparity between machine processing and human processing of emotions and abstract art, which necessitates further research from a cognitive modeling perspective.

Limitations include the subjective nature of abstract art, posing challenges for human annotators and CLIP. Social and cultural factors influence emotional judgments, hindering accurate insights into CLIP’s performance as human annotators also demonstrated disagreement in emotion labeling. To acknowledge emotion’s ambiguous nature, emotion classification should be operationalized as a multi-label task. The predominant focus on negative emotional states with a limited representation of positive emotions calls for future research to use emotion classes rooted in established psychological theories, such as Mikels et al. (2005)’s model. This could offer a more cognitively meaningful coverage of emotions to enhance AI model’s interpretability and relevance to human emotional experiences.

Acknowledgments

We thank Raquel Fernández and Esam Ghaleb for their feedback on a preliminary version of this work, and the anonymous CogSci reviewers for their valuable comments on the paper. ET worked on this project while she was employed at the University of Amsterdam, where this research received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 819455).

References

Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., & Brundage, M. (2021). Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

Ananthram, A., Winn, O., & Muresan, S. (2023). Feelingblue: A corpus for understanding the emotional connotation of color in context. *Transactions of the Association for Computational Linguistics*, *11*, 176–190.

Anderson, A. J., Zinszer, B. D., & Raizada, R. D. (2016). Representational similarity encoding for fmri: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, *128*, 44–53.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current directions in psychological science*, *20*(5), 286–290.

Beinborn, L., & Hollenstein, N. (2024). *Cognitive plausibility in natural language processing* (1st ed.). Springer Nature.

Berrios, W., Mittal, G., Thrush, T., Kiela, D., & Singh, A. (2023). Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*.

Bondielli, A., Passaro, L. C., et al. (2021). Leveraging clip for image emotion recognition. In *Ceur workshop proceedings* (Vol. 3015).

Casas, A., & Williams, N. W. (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, *72*(2), 360–375.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Flynn, T. N., & Marley, A. A. (2014). *Best-worst scaling: theory and methods*. Unpublished doctoral dissertation, Edward Elgar Worcester, UK.

Hemphill, M. (1996). A note on adults' color–emotion associations. *The Journal of genetic psychology*, *157*(3), 275–280.

Hess, U., & Hareli, S. (2016). The impact of context on the perception of emotions. *The expression of emotion: Philosophical, psychological, and legal perspectives*, 199–218.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7514–7528).

Kennedy, W. G. (2009). Cognitive plausibility in cognitive modeling, artificial intelligence, and social simulation. In *Proceedings of the international conference on cognitive modeling (iccm), manchester, uk* (pp. 24–26).

Ko, E., Yoon, C., & Kim, E. Y. (2016). Discovering visual features for recognizing user's sentiments in social images. In *2016 international conference on big data and smart computing (bigcomp)* (pp. 378–381).

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, 23–29 Jul). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 19730–19742). PMLR.

Liu, Y. (2022). The colour-emotion association. *Journal of Education, Humanities and Social Sciences*, *5*, 272–277.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the acl-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics* (pp. 63–70).

Lu, X., Suryanarayan, P., Adams Jr, R. B., Li, J., Newman, M. G., & Wang, J. Z. (2012). On shape and the computability of emotions. In *Proceedings of the 20th acm international conference on multimedia* (pp. 229–238).

Mesquita, B., & Boiger, M. (2014). Emotions in context: A sociodynamic model of emotions. *Emotion Review*, *6*(4), 298–302.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., & Reuter-Lorenz, P. A. (2005). Emotional category data on images from the international affective picture system. *Behavior research methods*, *37*, 626–630.

Mittal, T., Bera, A., & Manocha, D. (2021). Multimodal and context-aware emotion perception model with multiplicative fusion. *IEEE MultiMedia*, *28*(2), 67–75.

Mohammad, S., & Kiritchenko, S. (2018). Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).*

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436–465.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on knowledge capture* (pp. 70–77).

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive science*, *39*(8), 1824–1854.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with clip latents.*

Sartori, A., Culibrk, D., Yan, Y., & Sebe, N. (2015). Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings. In *Proceedings of the 23rd acm international conference on multimedia* (pp. 311–320).

Sartori, A., Yanulevskaya, V., Salah, A. A., Uijlings, J., Bruni, E., & Sebe, N. (2015). Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *5*(2), 1–27.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., ... Keutzer, K. (2022). How much can CLIP benefit vision-and-language tasks? In *International conference on learning representations.*

Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A., & Shah, A. (2018). Review on natural language processing (nlp) and its toolkits for opinion mining and sentiment analysis. In *2018 ieee 5th international conference on engineering technologies and applied sciences (icetas)* (pp. 1–4).

Sutton, T. M., & Altarriba, J. (2016). Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior research methods*, *48*, 686–728.

Yang, D., Huang, S., Wang, S., Liu, Y., Zhai, P., Su, L., ... Zhang, L. (2022). Emotion recognition for multiple context awareness. In *European conference on computer vision* (pp. 144–162).

Yanulevskaya, V., van Gemert, J. C., Roth, K., Herbold, A.-K., Sebe, N., & Geusebroek, J.-M. (2008). Emotional valence categorization using holistic image features. In *2008 15th ieee international conference on image processing* (pp. 101–104).

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., & Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd acm international conference on multimedia* (pp. 47–56).