Annotating Aesthetics

Céline Offerman, Willem van der Maden, Alessandro Bozzon

Delft University of Technology
Delft, Netherlands
c.e.offerman@tudelft.nl, w.l.a.vandermaden@tudelft.nl, a.bozzon@tudelft.nl

Abstract

Recent advancements in text-to-image models can generate high quality images from prompts, but lack diversity for vague prompts like "beautiful". This may stem from limitations in training data. The LAION-Aesthetics dataset was constructed by rating images for liking. However, philosophical and empirical aesthetics research indicates that aesthetics involves appraisal and contemplation beyond liking. This study tested four hypotheses to study alternative aesthetics annotation methods for images: activating semantic concepts, rating aesthetic value, ranking images, and 2-alternative forced choice preference. Experiments with crowdworkers found no alternative which outperformed the baseline. Further research can look into diverse image classes, consider regional/demographic influences, annotator expertise, and vary annotation contexts.

Introduction

Recent advancements in text-to-image models (a form of Generative Artificial Intelligence), such as *Stable Diffusion*[1], enable the generation of high quality images from textual prompts. However, we observe limitations in the diversity of outputs when using vague prompts like "beautiful". As seen in Figure 1, images generated from this prompt predominately depict women and flowers. This is despite the term "beautiful" encompassing a wide range of aesthetics across people, objects, scenes. In this work-in-progress, we ask, how is this possible? There may be different reasons why this observation occurs. One plausible explanation could lie within the training dataset of Stable Diffusion.

Figure 1: Images generated with Stable Diffusion, using the prompt *"beautiful"*.

Presented at the Works-in-Progress and Demonstrations track, AAAI HCOMP 2023.

[1] www.stability.ai

SD was trained on the LAION-Aesthetics 5+ dataset[2], a large collection of images and associated alt-text, controlled for aesthetic value of the images (Schuhmann 2022). This dataset was constructed by asking the simple question *"How much do you like this image on a scale from 1-10?"*. As will be described in the related work section, an aesthetic experience is a complex sensory experience. The process of inquiring participants about their image *liking* appears to not fully encompass the intricacies of this multifaceted human experience.

This paper investigates the challenge of AI-generated images lacking diverse aesthetic topics related to specific prompts, despite minimal explicit guidance during generation. We do so by following different lines of inquiry, experimenting with different questions and modalities. In the future, it is anticipated that crowdworkers will be involved to annotate these datasets for models like Stable Diffusion.

Related work

This section addresses related literature concerned with aesthetics, and highlights the two lines of inquiry followed in this project, namely the impact of question and response framing and effects of response modalities.

Aesthetics in philosophy

From a philosophical and neuropsychological perspective, aesthetics is considered to be a sensory experience (Baumgarten 2007; Merleau-Ponty 2011; Mandoki 2007; Saito 2010), which leads to a disinterested encounter (Kant 2000), involving appraisal (Simpson 1975) and inducing a contemplative state (Schopenhauer 2010; Chatterjee 2002, 2003; Vartanian and Skov 2014) in beholders.

Construction LAION-5b

To create the training dataset from Stable Diffusion, a predictor was trained that ascribed an 'aesthetic' score to images. This predictor was deployed to bucket the photos from the LAION 5B dataset for aesthetic scores. All images with an ascribed aesthetic score of 5+ were included in the training dataset (Schuhmann 2022). This predictor is thus respon-

[2] available at https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_6.5plus

sible for including and excluding images in Stable Diffusion's training dataset.

Schuhmann (2022) describes that to create the predictor, several models were trained that predict the rating people gave images when they were asked *"How much do you like this image on a scale from 1 to 10?"*. This annotation instruction is not backed by literature to actually measure aesthetics. With the theory described in the previous subsection in mind, the procedure of asking participants about image *liking* does not seem to cover the load of this complex human experience.

Impact of question and response framing

There are numerous ways you can have participants answer questions, and this affects the answers these participants provide. Although the comparison of different modalities may seem unconventional, it has been done in existing literature, e.g. different question formulations (Semin and De Poot 1997b), rating vs. ranking (Alwin and Krosnick 1985), rating vs. 2 alternative forced choice (Yannakakis and Hallam 2011).

The Unified Model of Aesthetics (UMA) integrates perceptual, cognitive, and social factors, including unity-in-variety, typicality & novelty, and connectedness & autonomy (Berghman and Hekkert 2017). Faerber et al. (2010) found that exposing participants to semantic concepts linked to aesthetic appreciation, significantly impacts their aesthetic experiences. This exposure affects subsequent ratings, demonstrating the link between semantic concept activation and aesthetic perception. This leads to the following hypothesis **[H1]**: exposing participants to semantic concept activation through questions about unity-in-variety, typicality & novelty, and connectedness & autonomy, will influence their ratings of how much they like images.

The way in which a question is phrased influences participant responses (Semin and De Poot 1997a,b). Various studies directly inquire about the aesthetic value of stimuli (Bhattacharya, Sukthankar, and Shah 2010; Datta et al. 2006; Zhang et al. 2014; Redi et al. 2013; J-D 2022). This literature forms the basis for the following hypothesis **[H2]**: The manner in which participants are asked a question on aesthetics significantly impacts their responses.

Effects of response modalities

Ranking, demonstrated as an effective modality (Nguyen et al. 2012; Yuan et al. 2023), is combined here with asking directly for aesthetic value, as described above, leading to the following hypothesis **[H3]**: Participants' rankings of image aesthetics will show significant differences when compared to their subjective ratings of image liking on a scale of 1-10.

The 2AFC (alternative forced choice) modality is valuable for assessing relative aesthetic preferences (Palmer, Schloss, and Sammartino 2013; Wu et al. 2023; Swanson, Escoffery, and Jhala 2012; Bara, Binney, and Ramsey 2021; Bıyık et al. 2020; Sadigh et al. 2017). This forms the basis of the following hypothesis **[H4]**: The alternative forced choice annotations of image aesthetics will result in significantly different outcome scores compared to participants' subjective ratings of image liking on a scale of 1-10.

Methods

Participants

Surveys were conducted on Prolific, aiming for n=10 per set of 10 stimuli. Crowdworkers were excluded from analysis for failing attention checks or for not completing surveys. Each treatment had $n = 20$ participants, with exceptions for the aesthetic value treatment ($n = 17$) and ranking treatment ($n = 19$).

Materials

Participants viewed 10 images with "building" alt-text from LAION 5B, rescaled to uniform size. The exploratory studies performed here aimed to investigate rather than validate, so results should be interpreted cautiously. Stimulus sets were controlled to have similar aesthetic distribution through collaboration between researchers and surveys ($n = 12$, $n = 60$) with design master students.

Procedure

Participants were invited via Prolific [3]. Each experiment began with training tasks that aligned with the respective treatment, as suggested by (Daniel et al. 2019). In each experiment, the control treatment was compared with an alternative treatment. In the control treatment, participants were asked to rate the stimuli on the following scale: *"how much do you like this image on a scale from 1 to 10?"*, as described by Schuhmann (2022).

- **The semantic concept activation treatment (Experiment 1):** In this treatment, participants encountered 12 items related to aesthetic appreciation, adapted from prior research (Berghman and Hekkert 2017). The social level of the UMA is interpreted as relatedness to make it context appropriate (Deci and Ryan 2000). Each item was rated on a scale from 1 to 10 and served as semantic concepts. After this exposure, participants were then asked to rate the stimulus for liking on the same 1 to 10 scale.
- **The aesthetic value treatment (Experiment 2):** Participants were instructed to assess the aesthetic value of images by responding to the question *"how aesthetic do you find this image?"* using a scale from 1 to 10.
- **The ranking treatment (Experiment 3):** For this treatment, participants were tasked to rank 10 stimuli based on their aesthetic value.
- **The 2AFC treatment (Experiment 4):** Here, participants were presented with pairs of images and asked to make forced-choice preference selections.

The experiments compare the control treatments to alternative treatments based on two main criteria: the alternative treatments yield significantly distinct results from the control treatment, and the internal consistency meets the acceptable threshold of Cronbach's alpha value ≥ 0.7. The analysis

[3] www.prolific.co

for these experiments were performed using JMP 17 software.

Results

H1: Semantic Concept Activation T-tests revealed no significant differences ($p \geq 0.12$) per image between treatments. Cronbach's alpha indicated satisfactory inter-rater reliability for both semantic concept activation treatment ($\alpha = 0.89$; $\alpha = 0.82$) and control treatment ($\alpha = 0.94$; $\alpha = 0.76$). This suggests that semantic concept activation has no significant impact on participants' liking ratings.

H2: Aesthetic Value T-tests showed no significant differences ($p \geq 0.07$) per image between aesthetic value treatment and control treatment. Inter-rater reliability comparisons indicated no systematic increase for the aesthetic value treatment ($\alpha = 0.84$; $\alpha = 0.93$) compared to the control treatment ($\alpha = 0.94$; $\alpha = 0.76$). This implies that rating images for aesthetic value and image liking are equivalently appropriate.

H3: Ranking Experiment 3 compared ranking and rating modalities. Linear regression analyses assessing the relationship between calculated mean scores and relative ranks demonstrated weak correlations ($slope = 0.13; slope = 0.27$), lacking statistical significance ($p = 0.38; 0.26$) with unacceptable R-squared values ($R-squared = 0.04; 0.07$). Cronbach's alpha indicated high reliability for control treatment ($\alpha = 0.78$; $\alpha = 0.85$) and poor reliability for ranking treatment ($\alpha = 0.41$; $\alpha = -0.82$). A negative Cronbach's alpha value is noteworthy. Having reviewed the data, it seems that this may be due to the small sample size. For this data, ranking images on aesthetic value performs significantly worse for the pre-set criteria.

H4: 2AFC Experiment 4 compared 2AFC and rating modalities. To compare the two, we rescaled the number of preferred stimulus clicks per participant to 1-10. No significant difference ($p = 0.75$) was found in mean scores per image. Cronbach's alpha indicated no increase in reliability for 2AFC treatment ($\alpha = 0.82$; $\alpha = 0.92$) compared to control treatment ($\alpha = 0.78$; $\alpha = 0.85$). This suggests that preference indication and image liking ratings are equivalently appropriate.

Discussion

Experiments found no significant differences between image liking ratings and alternative aesthetics annotation methods like concept activation, aesthetic value ratings, ranking, or 2AFC for this stimulus set. Despite hypotheses that alternate techniques may better capture aesthetics, none clearly improved upon liking ratings in terms of distinctiveness or reliability. Based on the literature discussed, these findings are surprising.

In fact, one of the alternative approaches, the ranking method, even performed significantly worse. It remains unclear whether the lack in diversity found in the predicted images is attributable to the question itself (*"How much do you like this image on a scale from 1-10?"*) or the manner in which the question was asked. To further explore this topic, it would be interesting to examine different image classes, including non-functional and/or controversial topics. In Kant's view, "free beauty" arises from appreciating the stimulus itself, while "dependent beauty" relates to how effectively the stimulus serves its intended purpose (Kant 2000). Functional stimuli might predominantly evoke a sense of dependent beauty. In this state of functional contemplation the viewer places value on the stimulus's practical attributes, rather than its aesthetic qualities in aesthetic contemplation. Furthermore, it is interesting to look at how non-aesthetic properties influence aesthetic scores. Thus, it is possible to look at how certain aspects of the stimuli (e.g. scariness) as confounding variables potentially influence participants' ascribed aesthetic scores.

Our post-hoc analysis did not reveal significant effects of region. However, prior work demonstrates aesthetic preferences vary across regions and time periods (Hekkert and Leder 2008; Berghman and Hekkert 2017). Further investigating regional influences remains important. If regions impact aesthetic judgments, this could raise questions around the potential need for developing region-specific training models.

Next to this, it is worth considering that crowdsourcing might not be the most appropriate method for measuring aesthetic experiences. Several studies suggest participants have heightened aesthetic experiences within museums versus lab contexts (Brieber, Nadal, and Leder 2015; Locher, Smith, and Smith 1999, 2001). In other words, investigating diverse annotation settings could provide valuable insights. We recommend that future research should involve diverse participant demographics to annotate images and compare with crowdworkers, and consider comparisons with aesthetic experts (e.g., Hosu et al. (2019) uses photographers as experts for image aesthetics).

Lastly, these experiments were designed primarily for exploration rather than validation, utilising small stimulus and sample sizes. Furthermore, the high inter-rater reliability observed in the control treatment might be influenced by the specific image class used, potentially limiting the applicability of our results. Future research should consider alternative image classes. To further explore the generalisability of our findings, additional research should investigate performance across diverse image classes, including natural scenes, abstract art, and provocative content. Stimuli specifically selected to evoke a spectrum of aesthetic reactions could better discriminate between annotation approaches.

Conclusion

In conclusion, this work-in-progress explored alternative methods for assessing aesthetic preferences to label image datasets that can be used for training Generative AI models. Despite formulating and testing various approaches, including semantic concept activation, aesthetic value rating, ranking, and 2AFC, none outperformed the LAION Aesthetics approach. Further research into different image classes, the influence of region, annotation contexts, and involving participants from various demographic groups are recommended for a comprehensive understanding of aesthetic experiences in this context.

References

Alwin, D. F.; and Krosnick, J. A. 1985. The Measurement of Values in Surveys: A Comparison of Ratings and Rankings. *The Public Opinion Quarterly*, 49(4): 535–552.

Bara, I.; Binney, R. J.; and Ramsey, R. 2021. Investigating the Role of Executive Resources across Aesthetic and Non-Aesthetic Judgments. preprint, PsyArXiv.

Baumgarten, A. G. 2007. esthetica/Ästhetik. 2.

Berghman, M.; and Hekkert, P. 2017. Towards a unified model of aesthetic pleasure in design. *New Ideas in Psychology*, 47: 136–144.

Bhattacharya, S.; Sukthankar, R.; and Shah, M. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM international conference on Multimedia*, 271–280. Firenze Italy: ACM. ISBN 9781605589336.

Brieber, D.; Nadal, M.; and Leder, H. 2015. In the white cube: Museum context enhances the valuation and memory of art. *Acta Psychologica*, 154: 36–42.

Bıyık, E.; Huynh, N.; Kochenderfer, M. J.; and Sadigh, D. 2020. Active Preference-Based Gaussian Process Regression for Reward Learning.

Chatterjee, A. 2002. Universal and relative aesthetics: a framework from cognitive neuroscience. Takarazuka, Japan.

Chatterjee, A. 2003. Prospects for a cognitive neuroscience of visual aesthetics: (514602010-003).

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2019. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys*, 51(1): 1–40.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In Leonardis, A.; Bischof, H.; and Pinz, A., eds., *Computer Vision – ECCV 2006*, Lecture Notes in Computer Science, 288–301. Berlin, Heidelberg: Springer. ISBN 9783540338376.

Deci, E. L.; and Ryan, R. M. 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, 11(4): 227–268.

Faerber, S. J.; Leder, H.; Gerger, G.; and Carbon, C.-C. 2010. Priming semantic concepts affects the dynamics of aesthetic appreciation. *Acta Psychologica*, 135(2): 191–200.

Hekkert, P.; and Leder, H. 2008. PRODUCT AESTHETICS. In *Product Experience*, 259–285. Elsevier. ISBN 9780080450896.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2019. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment.

J-D, P. 2022. Simulacra Bot.

Kant, I. 2000. *Critique of the Power of Judgment*. Cambridge University Press, 1 edition. ISBN 9780521348928 9780521344470 9780511804656.

Locher, P.; Smith, L.; and Smith, J. 1999. Original Paintings versus Slide and Computer Reproductions: A Comparison of Viewer Responses. *Empirical Studies of the Arts*, 17(2): 121–129.

Locher, P. J.; Smith, J. K.; and Smith, L. F. 2001. The Influence of Presentation Format and Viewer Training in the Visual Arts on the Perception of Pictorial and Aesthetic Qualities of Paintings. *Perception*, 30(4): 449–465.

Mandoki, K. 2007. *Everyday Aesthetics: Prosaics, the Play of Culture and Social Identities (2007, 2016 Routledge)*.

Merleau-Ponty, M. 2011. *Oog en geest*. Amsterdam: Parrèsia, heruitg. edition. ISBN 9789073040007. OCLC: 812537916.

Nguyen, T. V.; Liu, S.; Ni, B.; Tan, J.; Rui, Y.; and Yan, S. 2012. Sense beauty via face, dressing, and/or voice. In *Proceedings of the 20th ACM international conference on Multimedia*, 239–248. Nara Japan: ACM. ISBN 9781450310895.

Palmer, S. E.; Schloss, K. B.; and Sammartino, J. 2013. Visual Aesthetics and Human Preference. *Annual Review of Psychology*, 64(1): 77–107.

Redi, J. A.; Hoßfeld, T.; Korshunov, P.; Mazza, F.; Povoa, I.; and Keimel, C. 2013. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 29–34. Barcelona Spain: ACM. ISBN 9781450323963.

Sadigh, D.; Dragan, A.; Sastry, S.; and Seshia, S. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems XIII*. Robotics: Science and Systems Foundation. ISBN 9780992374730.

Saito, Y. 2010. *Everyday aesthetics*. Oxford ; New York: Oxford University Press. ISBN 9780199575671. OCLC: ocn441191695.

Schopenhauer, A. 2010. *The world as will and representation*. The Cambridge edition of the works of Schopenhauer. Cambridge ; New York: Cambridge University Press. ISBN 9780521871846 9780521870344.

Schuhmann, C. 2022. LAION-Aesthetics | LAION.

Semin, G. R.; and De Poot, C. J. 1997a. Bringing Partiality to Light: Question Wording and Choice as Indicators of Bias. *Social Cognition*, 15(2): 91–106.

Semin, G. R.; and De Poot, C. J. 1997b. The question–answer paradigm: You might regret not noticing how a question is worded. *Journal of Personality and Social Psychology*, 73(3): 472–480.

Simpson, E. 1975. Aesthetic Appraisal. *Philosophy*, 50(192): 189–204.

Swanson, R.; Escoffery, D.; and Jhala, A. 2012. 2012 IEEE Conference on Computational Intelligence and Games (CIG).

Vartanian, O.; and Skov, M. 2014. Neural correlates of viewing paintings: Evidence from a quantitative meta-analysis of functional magnetic resonance imaging data. *Brain and Cognition*, 87: 52–56.

Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; and Li, H. 2023. Better Aligning Text-to-Image Models with Human Preference.

Yannakakis, G. N.; and Hallam, J. 2011. Ranking vs. Preference: A Comparative Study of Self-reporting. In D'Mello,

S.; Graesser, A.; Schuller, B.; and Martin, J.-C., eds., *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, 437–446. Berlin, Heidelberg: Springer. ISBN 9783642246005.

Yuan, Z.; Yuan, H.; Tan, C.; Wang, W.; Huang, S.; and Huang, F. 2023. RRHF: Rank Responses to Align Language Models with Human Feedback without tears.

Zhang, T.; Nefs, H. T.; Redi, J.; and Heynderickx, I. 2014. The aesthetic appeal of depth of field in photographs. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 81–86. Singapore, Singapore: IEEE. ISBN 9781479965366.