Automatic Aesthetic Evaluation in Generative Image Models

Larissa D. Gomide, Lucas N. Ferreira, Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
{larissa.gomide, lferreira, meira}@dcc.ufmg.br

Abstract

Recent art assessment neural models have shown promising results as aesthetic evaluators of paintings, computing objective metrics through handcrafted features and deep learning techniques. While these models have been applied to evaluate human-produced pieces, their use in measuring the aesthetic quality of artificial intelligence (AI)-generated artistic works remains largely unexplored. This paper employs ArtCLIP, a recent art assessment neural model, to evaluate the aesthetic quality of paintings produced by state-of-the-art image generation models. Our methodology includes automatically describing human artworks with an image-to-text model and generating new images from these prompts, allowing us to pair human artworks with generated pieces. We compare the distribution of nine different aesthetic attributes given by ArtCLIP between human and AI-generated paintings. We considered models of two considerably different sizes to measure the impact of model size on aesthetic quality. The results show that ArtCLIP ranks human artwork as having higher aesthetic quality, with the larger generator outperforming the smaller one.

Introduction

In Artificial Intelligence, generative models are often evaluated using objective metrics such as the Bilingual Evaluation Understudy (BLEU) score for Machine Translation (Papineni et al. 2002), and the Fréchet Inception Distance (FID) for Computer Vision (Heusel et al. 2017) tasks. These metrics have played a central role in driving progress across domains by providing standardized, quantifiable benchmarks. However, such measures are primarily concerned with fidelity and realism, which—while important—are insufficient when evaluating outputs in Computational Creativity, where the goals extend beyond imitation toward novelty, value, and surprise (Colton and Wiggins 2012).

In Computational Creativity, human-centered assessments—including expert critique and user studies—remain the gold standard but are time-consuming and difficult to scale or reproduce. Thus, it is important to develop automated yet meaningful evaluation metrics that can approximate subjective judgment while being generalizable and consistent (Jordanous 2022).

In this direction, (Jin et al. 2024b) introduced the Aesthetics of Paintings and Drawings Dataset (APDDv2), which consists of 10,023 human artworks annotated with scores for ten artistic attributes frequently discussed in visual arts literature, such as mood, creativity, layout, and composition, among others. Alongside, they proposed ArtCLIP, a collection of predictive models trained to estimate these attributes. While ArtCLIP shows strong performance on human-authored pieces, its applicability to AI-generated visual art has not been examined—raising the question of whether such models can serve as robust and meaningful aesthetic evaluators in creative AI systems.

In this paper, we address this gap by evaluating whether ArtCLIP can function as an aesthetic scoring model for generated paintings. First, we built a dataset of generated paintings by employing the image-to-text Janus-Pro model (Chen et al. 2025) to describe a sample of human paintings from the APDDv2 dataset. Then, we used these textual descriptions to generate new paintings with two versions of the text-to-image Janus-Pro model: a small (≈1B parameters) and a large (≈7B parameters) version. This process yielded a dataset of 1,000 paired samples (I_h, I_{gs}, I_{gl}, T), where I_h is the original human-made artwork, I_{gs} and I_{gl} are paintings generated by the small and large models, respectively, and T is their shared textual description. Finally, we employed ArtCLIP to compute nine aesthetic attributes for each image independently.

To assess ArtCLIP as an objective aesthetic metric of visual artwork, we first compared the distributions of the nine ArtCLIP scores across the human, Janus-Pro-1B, and Janus-Pro-7B paintings. Results revealed that pieces by Janus-Pro-1B received lower scores than those from Janus-Pro-7B, which in turn were rated lower than the human-created artworks. We also evaluated the consistency of ArtCLIP across human, Janus-Pro-1B, and Janus-Pro-7B paintings—that is, whether paintings with the same description received similar scores. For each attribute, we computed the average distance between the scores assigned to human artworks and their generated counterparts. As expected, the distances between the human and the large model were consistently smaller than those between the human and the small model across all attributes. Overall, these findings suggest that ArtCLIP can serve as a consistent metric for assessing the aesthetic quality of AI-generated paintings.

The main contributions of this paper are:

- A dataset of human paintings from APDDv2 paired with two generated counterparts—one produced by a small model and the other by a large model—based on a shared textual description;
- An evaluation of ArtCLIP as an objective metric for assessing the aesthetic quality of paintings produced by image generation models.

Related Work

Our work is directly related to ArtCLIP and the APDDv2 dataset (Jin et al. 2024b), which are the primary focus of this study. It is also connected to previous efforts to develop objective metrics for evaluating creative systems. This section briefly reviews each of these topics.

APDDv2 and ArtCLIP

The APDDv2 dataset comprises 10,023 paintings across 24 artistic categories, with each image annotated according to 10 aesthetic attributes. These categories vary by painting type, artistic style, and subject matter—for example, *Oil Painting - Symbolism - Landscapes* and *Traditional Chinese Painting - Meticulous - Portraiture*. The aesthetic attributes include: *Theme and Logic* (T&L), *Creativity* (Cre), *Layout and Composition (L&C)*, *Space and Perspective* (S&P), *Sense of Order* (SO), *Light and Shadow* (L&S), *Color* (Col), *Details and Texture* (D&T), *Mood* (M), and *The Overall* (TO). Each image was evaluated on a continuous scale from 0 to 10 by at least six annotators. The annotation team consisted of 37 trained individuals, including professional artists and educators with at least a bachelor's degree and extensive experience.

Using this dataset, the authors introduced ArtCLIP, a multimodal aesthetic assessment model that combines image and text embeddings through contrastive learning. The model is initially pre-trained on the DPC2022 dataset (Zhong, Zhou, and Qiu 2023), which contains photographic images paired with categorized aesthetic comments, enabling ArtCLIP to learn general aesthetic representations. It is then fine-tuned on APDDv2 to specialize in evaluating paintings. Compared to baseline models AANSPS (Jin et al. 2024a) and SAAN (Yi et al. 2023), ArtCLIP demonstrates improved or comparable performance, showing higher agreement with human annotations and lower prediction error on attributes such as *Composition*, *Color*, and *Mood*.

Other related image datasets and aesthetic evaluation models also inform our work. For example, (Achlioptas et al. 2021) introduced ArtEmis, a dataset of artworks annotated with emotional labels and natural language explanations, which was used to train models capable of generating affective captions. Similarly, (Vera Nieto, Celona, and Fernandez Labrador 2022) presented the Reddit Photo Critique Dataset, which contains photographs paired with user-generated critiques and derives aesthetic scores through sentiment analysis. Both approaches rely on subjective, unstructured language to infer aesthetic judgments. In contrast, our work evaluates structured, expert-defined aesthetic attributes using ArtCLIP and applies this model to both human and AI-generated paintings.

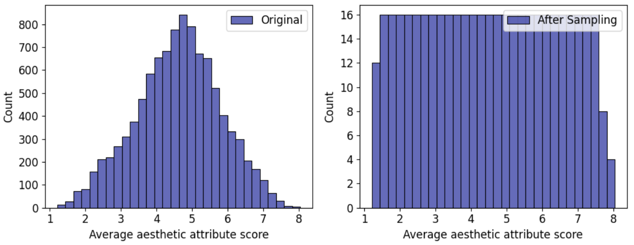


Figure 1: The original distribution of average aesthetic attribute scores divided into 30 bins (left), and the distribution after sampling (right).

Creativity Metrics

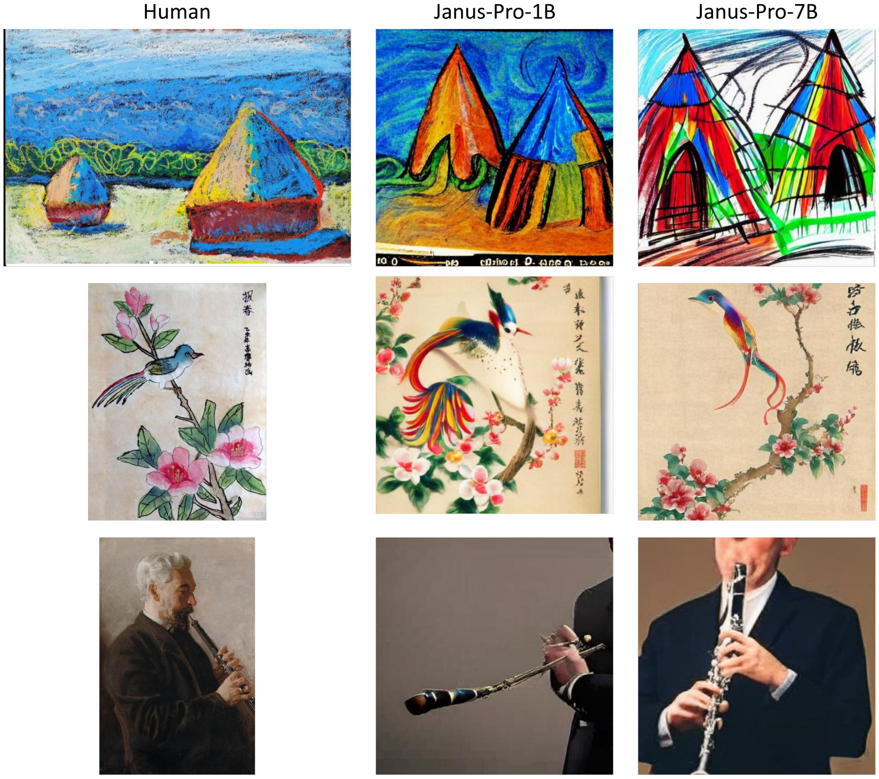
Our work is also related to ongoing efforts to quantify creativity in machine-generated content. For instance, (Wang et al. 2023) investigates how well existing metrics such as FID (Heusel et al. 2017) align with human judgments when evaluating combinational creativity in images generated by DALL-E (Ramesh et al. 2021). Another example is the work of (Lu et al. 2024), in which the authors propose the *Creativity Index*, a metric for assessing linguistic creativity by measuring how much of a given text can be reconstructed from existing web sources. The primary distinction between these approaches and ours is that, rather than evaluating creativity exclusively, our work explores nine different attributes of aesthetic quality of visual outputs, including creativity as one of them.

More broadly, this work contributes to the ongoing conversation about evaluation metrics in Computational Creativity, a field that has long grappled with the challenge of developing standardized, objective evaluation practices. As noted by (Jordanous 2022), the field has reached a point of maturity where establishing such metrics is increasingly feasible. Our study adds to this effort by investigating structured, attribute-based evaluation of visual aesthetics in AI-generated paintings.

Methodology

Our main goal is to evaluate ArtCLIP as an aesthetic scoring model for generated paintings. To do this, we created a new dataset of generated paintings paired with corresponding human-made versions from APDDv2. Since the original dataset contains approximately 10k images, generating counterparts for all of them using state-of-the-art image generators would be computationally expensive. To manage these costs while maintaining statistical validity, we employed a sampling strategy and selected 500 images from the original dataset. Based on the standard sample size formula using a z-score for confidence intervals, at least 385 instances are required to achieve 95% confidence with a 5% margin of error.

The original distribution of paintings in APDDv2 is not balanced in terms of average aesthetic attribute score. In



This image appears to be a colorful, abstract drawing or painting. It features two prominent, conical structures that resemble tents or huts, with vibrant colors such as blue, red, yellow, and orange. The background consists of a blue sky with some white and black swirling patterns, and a green, wavy line that could represent vegetation or a fence. The overall style is expressive and uses bold, contrasting colors to create a dynamic and lively scene.

The image is a traditional Chinese painting featuring a bird perched on a branch. The bird has a blue head, a white body, and a colorful tail with red, yellow, and green hues. It is surrounded by blooming flowers, primarily pink blossoms with green leaves. There is also some Chinese calligraphy on the right side of the painting.

The image depicts a person playing a clarinet. The individual is dressed in a dark suit and is holding the clarinet with both hands. The background is a plain, muted color, which helps to focus attention on the subject.

Figure 2: Four examples of instances from our dataset. An example is composed of a human-authored piece from APDDv2 paired with two images: one generated with Janus-Pro-1B and another with Janus-Pro-7B. The three images share the same textual description, which the image-to-text Janus-Pro-7B model gives.

particular, it follows a normal distribution, meaning there are considerably fewer images with maximum (10) and minimum (0) scores than with intermediate scores (around 5). To address this, we divided the images into 30 bins based on their average aesthetic scores and selected an equal number of samples from each bin, resulting in approximately 16 images per bin. Figure 1 shows the original distribution (left) and the balanced distribution after sampling (right). This strategy ensures that ArtCLIP's performance can be evaluated consistently across the full range of aesthetic scores, including extreme cases.

With this balanced sample of 500 paintings, we used an image-to-text model to produce textual descriptions for each image, which we then used as prompts to generate new images with a text-to-image model. Specifically, we employed Janus-Pro-7B (Chen et al. 2025) for generating the descriptions. We chose Janus-Pro-7B because it is a state-of-the-art model that unifies multimodal understanding and generation. This allowed us to use it both for generating textual descriptions of the APDDv2 paintings and for creating new paintings from those descriptions. The smaller model tends to produce images that are closer to the original, effectively capturing the core content from the description—although it occasionally introduces small artifacts (e.g., a black border at the bottom). The larger model introduces more stylistic variation, diverging in tone and interpretation.

Each of the 500 sampled images was passed through Janus-Pro-7B with the prompt "Describe the image" to obtain its textual description. On average, the descriptions contained 88.31 words, with a standard deviation of 25.60. The longest description contained 262 words, and the shortest 38 words. Although Janus-Pro-7B supports direct image-to-image operations due to its multimodal nature, we opted to use textual descriptions as intermediate representations. This choice improves explainability, enables prompt-level analysis, and allows us to better understand how image generation is influenced by descriptive content.

Next, for each of the 500 descriptions, we generated two new paintings: one using Janus-Pro-1B and another using Janus-Pro-7B. These models share the same architecture but differ in size—the former has approximately 1 billion parameters, while the latter has around 7 billion. Each image was generated using the exact text description as a prompt, with no additional augmentation. At the end of this process, our dataset contained 1,500 images: 500 human-authored pieces from APDDv2, each paired with two AI-generated counterparts.

Figure 2 presents three examples from our dataset, each featuring a different painting type, artistic style, and subject matter. These examples show that the image-to-text model can describe not only the subject matter of the images but also stylistic elements such as color and painting type (e.g., abstract vs. traditional Chinese). Moreover, the text-to-image models appear to better capture the subject matter in abstract paintings (e.g., the two abstract huts) than in more realistic ones (e.g., the Chinese bird and the person playing the clarinet).

Finally, we applied the ArtCLIP scoring model to obtain aesthetic scores for the 1,500 images. ArtCLIP consists of 10 independent models, each corresponding to a different aesthetic attribute. During our evaluation, we couldn't run the model responsible for the "Sense of Order" attribute due to technical issues, and hence it was excluded. Therefore, we report results on 9 aesthetic scores.

To execute the steps of generating descriptions, images, and scores, which involve inferences with complex deep learning-based models, we used a high-performance computer equipped with a Ryzen 9 5950X 16-core processor, 64 GB of RAM, and an RTX 3090 Ti 24 GB graphics card.

Experiments and Results

We conducted two experiments to evaluate whether ArtCLIP can serve as a consistent and meaningful aesthetic evaluation model for AI-generated paintings. The first experiment compared the distributions of aesthetic attribute scores between the Human, Janus-Pro-1B, and Janus-Pro-7B artworks. The second experiment analyzed the consistency of ArtCLIP's judgments across these attributes, specifically, whether artworks derived from the same prompt (textual description) receive comparable scores.

Comparison of Aesthetic Scores

The goal of the first experiment was to assess whether ArtCLIP can distinguish between varying levels of aesthetic quality in a meaningful and interpretable way. By comparing the distributions of aesthetic scores assigned to Human, Janus-Pro-1B, and Janus-Pro-7B, we evaluated whether ArtCLIP's outputs reflect expected qualitative differences. If ArtCLIP is a reliable aesthetic evaluator, it should assign higher scores to originally produced human-created artworks. In contrast, generated images should receive slightly lower scores, since they are transformations of human ideas that may accumulate noise or inconsistencies across the generation, description, and evaluation pipeline, especially for smaller models.

Table 1 shows the nine mean aesthetic scores for Human, Janus-Pro-1B, and Janus-Pro-7B paintings, as given by ArtCLIP. For each attribute, we ran a Friedman test across painting sources and, if significant, a pairwise Wilcoxon test with Bonferroni correction. Based on the Friedman test, all nine criteria showed significant differences across the methods (p-value ≤ 0.05). The post-hoc pairwise Wilcoxon test showed that, for all attributes, the human paintings were significantly better than the Janus models, and the large model was significantly better than the small one. Therefore, we can conclude that human artworks consistently received the highest scores, followed by those from the large model, and lastly from the small model.

Table 1: Mean ± standard deviation of each aesthetic attribute (Attr.) for Human, Janus-Pro-1B, and Janus-Pro-7B paintings. For each attribute, the higher the score, the better. Different superscript letters indicate statistically significant differences ($p < 0.05$). A value in bold is the best score for a particular attribute (row).

Attr.	**Human**	**Janus-1B**	**Janus-7B**
T&L	**6.25**^{a} ± 1.23	5.89^{b} ± 0.86	6.14^{c} ± 0.82
Cre	**5.91**^{a} ± 1.41	5.65^{b} ± 0.81	5.82^{c} ± 0.82
L&C	**6.21**^{a} ± 1.35	5.69^{b} ± 1.06	6.01^{c} ± 1.01
S&P	**6.37**^{a} ± 1.38	5.74^{b} ± 1.06	6.09^{c} ± 1.04
L&S	**6.29**^{a} ± 1.45	5.53^{b} ± 1.23	5.94^{c} ± 1.21
Col	**6.19**^{a} ± 1.38	6.04^{b} ± 0.98	6.20^{c} ± 1.00
D&T	**6.10**^{a} ± 1.52	5.33^{b} ± 1.25	5.72^{c} ± 1.23
TO	**6.26**^{a} ± 1.41	5.76^{b} ± 1.06	6.09^{c} ± 1.02
M	**6.03**^{a} ± 1.35	5.53^{b} ± 1.00	5.85^{c} ± 1.01

These results suggest that ArtCLIP can capture a significant aesthetic gap between AI-generated and human-created paintings. Moreover, the ranking—Human $>$ Janus-Pro-7B $>$ Janus-Pro-1B across all attributes implies that increasing model capacity improves aesthetic performance, but not enough to match human-level quality.

Consistency

In the second experiment, we evaluated ArtCLIP's consistency by measuring how closely the aesthetic attribute scores of generated images aligned with those of their corresponding human counterparts. For each aesthetic attribute, we computed the average absolute difference between the scores assigned to each human artwork and its small and large model versions.

As shown in Figure 3, the large model yields smaller differences compared to the small model across all attributes, indicating closer alignment with human aesthetic judgments. The differences between the models' average absolute distances typically fall within a range of approximately 0.3 to 0.4, suggesting that the improvement gained from scaling up the model is relatively uniform across aesthetic dimensions. This result indicates that increasing model capacity benefits all aspects of aesthetic alignment to a similar degree, rather than disproportionately enhancing performance on specific attributes. Notably, the large model achieves an almost perfect match in the Color attribute, implying that Janus-Pro-7B is capable of replicating human-like color use with high fidelity, at least when prompted with automatically generated descriptions of images from the APDDv2 dataset.

Overall, these results suggest that ArtCLIP not only detects aesthetic differences between human and AI-generated paintings but also produces consistent scores when comparing artworks derived from the same description. This consistency across paired samples provides further evidence that ArtCLIP can function as a reliable metric for aesthetic evaluation in generative image systems.

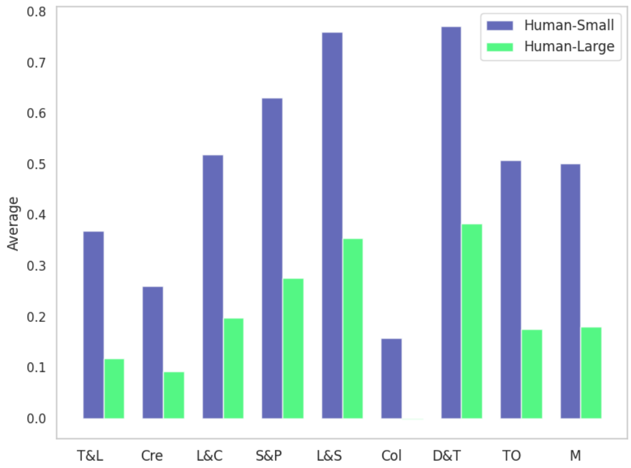


Figure 3: Average absolute differences in aesthetic attribute scores between original human artworks and their AI-generated counterparts.

Conclusion

This paper explored the use of ArtCLIP, a recent art assessment neural model, as an automatic evaluator of aesthetic quality in AI-generated paintings. To this end, we created a dataset of 500 human paintings, each paired with two generated images—one produced by a smaller Janus-Pro model (≈1B parameters) and another by a larger Janus-Pro model (≈7B parameters). All images share the same textual descriptions, generated using an image-to-text Janus-Pro model. We conducted two experiments to assess the reliability and consistency of ArtCLIP's aesthetic judgments across its nine aesthetic attribute scores.

Our results show that ArtCLIP effectively differentiates between human and machine-generated art, with human pieces receiving consistently higher scores across all nine aesthetic dimensions. Notably, we observed that larger generative models produce artwork that aligns more closely with human standards, both in overall score distributions and in attribute-level similarity. Furthermore, ArtCLIP demonstrated consistency in evaluating visually and semantically related images, reinforcing its potential as a scalable aesthetic metric. These findings suggest that ArtCLIP can serve as a useful proxy for subjective human judgment in aesthetic evaluation, offering a promising direction for automated benchmarking in creative AI systems.

Future work could explore how ArtCLIP ranks artwork from artists belonging to different art movements. Qualitative user studies may also help validate ArtCLIP's predictions against human perception. Another promising direction is to develop image generators guided explicitly by the nine ArtCLIP attributes.

Acknowledgments

This work is partially supported by the CNPq, Capes, Fapemig agencies, and the IAIA-Inct on AI project.

References

Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; and Guibas, L. J. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11569–11579.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *ECAI 2012*. IOS Press. 21–26.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30.

Jin, X.; Qiao, Q.; Lu, Y.; Gao, S.; Huang, H.; and Li, G. 2024a. Paintings and drawings aesthetics assessment with rich attributes for various artistic categories. *arXiv preprint arXiv:2405.02982*.

Jin, X.; Qiao, Q.; Lu, Y.; Wang, H.; Huang, H.; Gao, S.; Liu, J.; and Li, R. 2024b. Apddv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments. *arXiv preprint arXiv:2411.08545*.

Jordanous, A. 2022. Should we pursue sota in computational creativity?

Lu, X.; Sclar, M.; Hallinan, S.; Mireshghallah, N.; Liu, J.; Han, S.; Ettinger, A.; Jiang, L.; Chandu, K.; Dziri, N.; et al. 2024. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Vera Nieto, D.; Celona, L.; and Fernandez Labrador, C. 2022. Understanding aesthetics with language: A photo critique dataset for aesthetic assessment. *Advances in Neural Information Processing Systems* 35:34148–34161.

Wang, B.; Zhu, Y.; Chen, L.; Liu, J.; Sun, L.; and Childs, P. 2023. A study of the evaluation metrics for generative images containing combinational creativity. *AI EDAM* 37:e11.

Yi, R.; Tian, H.; Gu, Z.; Lai, Y.-K.; and Rosin, P. L. 2023. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22388–22397.

Zhong, Z.; Zhou, F.; and Qiu, G. 2023. Aesthetically relevant image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3733–3741.