Language Does More Than Describe: On The Lack Of Figurative Speech in Text-To-Image Models

Ricardo Kleinlein **Cristina Luna-Jiménez**

Fernando Fernández-Martínez
Grupo de Tecnología del Habla y Aprendizaje Automático
Information Processing and Telecommunications Center
E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid 28040 Madrid, Spain
{ricardo.kleinlein, cristina.lunaj, fernando.fernandezm}@upm.es

Abstract

The impressive capacity shown by recent text-to-image diffusion models to generate high-quality pictures from textual input prompts has leveraged the debate about the very definition of art. Nonetheless, these models have been trained using text data collected from content-based labelling protocols that focus on describing the items and actions in an image, but neglect any subjective appraisal. Consequently, these automatic systems need rigorous descriptions of the elements and the pictorial style of the image to be generated, otherwise failing to deliver. As potential indicators of the actual artistic capabilities of current generative models we characterise the sentimentality, objectiveness and degree of abstraction of publicly available text data used to train current text-to-image diffusion models. Considering the sharp difference observed between their language style and that typically employed in artistic contexts, we suggest generative models should incorporate additional sources of subjective information in their training in order to overcome (or at least to alleviate) some of their current limitations, thus effectively unleashing a truly artistic and creative generation.

1 The importance of figurative speech in art

Deep Generative Models (DGM) have become a particularly hot research topic within the field of machine learning [22]. DGMs are neural networks with thousands of parameters trained to approximate probability distributions from the observation of a large number of samples, after which they can be used to create new instances that resemble the training data. Their popularity has been steadily increasing in the last years, reaching the public sphere due to the success of autoregressive models [31, 35], Variational AutoEncoders (VAEs) [15, 16] and Generative Adversarial Networks (GANs) [11]. However, to condition the generation process towards more elaborate image semantics has been a long-standing problem [12, 8, 20, 27, 23]. With the advent of Large Language Models (LLMs) [25, 5] and CLIP [24], diffusion models can now be conditioned by textual prompts in natural language [13]. Still, some of the most popular ones such as DALL-E2 [26], Stable Diffusion [28] and Imagen [30] need these prompts to describe as exactly as possible both the items and the pictorial style of the image to generate.

It is not a minor issue, given how tightly bounded human communication and art are. Even in indigenous cultures [4], art is used to hand information about the feelings, ideas and perception of the artist to an audience, arousing emotions along the process [34, 19]. People often describe artworks in terms associated with emotional states [7], and as some theorists of art have noted, the exact content of a visual artwork is neither arbitrary nor unimportant, but perfectly aligned with the abstract theme

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

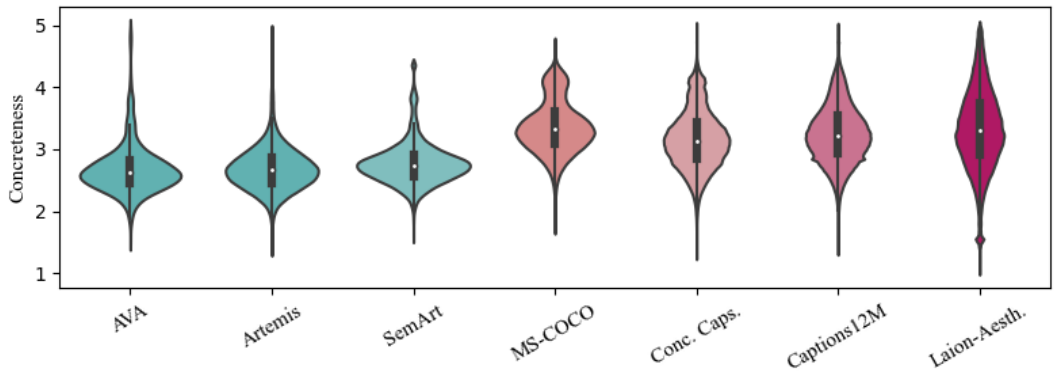

Figure 1: Distributions of average sentence concreteness score in each dataset. The color flips represent the shape of the distribution whereas the black box inside denotes the mean value and standard deviation of the distribution.

the artist seeks to communicate [2]. Furthermore, the elements in an artwork do not even need to be realistic to achieve its goal [3]. We argue current text-to-image generative models are ignoring this fundamental aspect of the human creative process, hence failing to deliver when the prompts conditioning the generation process of the model is not strictly descriptive (Appendix A).

2 Experimental setup and results

Our main argument is that text-to-image generative models are being trained on databases originally created to different purposes like image segmentation, hence limiting the way they interpret human language in the specific context of art generation. SemArt [10] and Artemis [1] are examples of datasets specifically aimed at understanding the role played by language in the description of artworks, thus paramount instances of the idiosyncrasy of the language used in artistic contexts. Similarly, incorporating to our study the descriptions of the challenges contained in the AVA dataset [21], we gain insight into how people ask for new pieces of art (in this case, photography contests). Opposed to them, MS-COCO [17], Conceptual Captions [33], Captions12M [9] and Laion-Aesthetics [32] are publicly available databases that constitute part of the training material used in text-to-image diffusion models whose language is essentially content-based and aimed to a neutral description of the scene. We qualitatively describe the sentiment valence [14], subjectivity [18] and average term concreteness [6] (the inverse of term abstraction) of image captions following the procedure introduced in [1].

The most obvious difference can be seen in the average term concreteness per sentence, shown in Figure 1. Apparently, datasets focused on art display more abstract vocabulary. When we consider their average sentiment valence, they also seem to be shifted towards positive emotions ($\mu = 0.27, \sigma^2 = 0.45$) while content-based data is fundamentally neutral ($\mu = 0.1, \sigma^2 = 0.27$). Likewise, Artemis, SemArt and AVA present on average a greater ratio of subjective terms ($\mu = 0.41, \sigma^2 = 0.27$) than the rest of datasets ($\mu = 0.22, \sigma^2 = 0.28$). We also compute the Earth-Moving Distance between each pair of datasets [29], finding that there are in fact two types of language styles, namely that related to art environments and that of content-based, neutral and objective descriptions (we refer the reader to Appendix B for further detail).

Acknowledgments and Disclosure of Funding

The work leading to these results was supported by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22, funded by MCIN/AEI/10.13039/501100011033), and AMIC-PoC (PDC2021-120846-C42, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR"). Ricardo Kleinlein's research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225).

3 Ethical considerations

The automatic generation of content, be it text, audio or image is now in the middle of a heated debate that may remind us of that happening decades ago when the massive adoption of photography was followed by the discussion about whether it could be considered an artistic expression [36]. Contrarily to cameras, the black-box nature of DGMs makes particularly challenging - although no less important - to understand the impact of the training material of the model on the final outcome to ensure fairness and prevent negative biases. We hope our work contributes to the demystification of these generative models while planting the seed for future work on their interpretability.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. *CoRR*, abs/2101.07396, 2021.

[2] R. Arnheim. *Art and Visual Perception*. Faber paper covered editions. University of California, 1954.

[3] C. Bell and J.B. Bullen. *Art*. A Grey Arrow book. Oxford University Press, 1987.

[4] Steven Brown and Ellen Dissanayake. The synthesis of the arts: From ceremonial ritual to "total work of art". *Frontiers in Sociology*, 3:9, 5 2018.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020-December, 5 2020.

[6] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911, 10 2014.

[7] N. Carroll. *Art in Three Dimensions*. OUP Oxford, 2010.

[8] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michał Drożdżal, and Adriana Romero-Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 33:27517–27529, 9 2021.

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3557–3567, 2 2021.

[10] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Understanding art through multi-modal retrieval in paintings. 4 2019.

[11] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. 2016.

[12] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. 5 2019.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020-December, 6 2020.

[14] C J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–225, 5 2014.

[15] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 12 2013.

[16] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12:307–392, 6 2019.

[17] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. volume 8693 LNCS, pages 740–755. Springer Verlag, 2014.

[18] Steven Loria. Textblob. Accessed September 16, 2022). Available at https://textblob.readthedocs.io/en/dev.

[19] Stefano Mastandrea, Sabrina Fagioli, and Valeria Biasi. Art and psychological well-being: Linking the brain to the aesthetic emotion. *Frontiers in Psychology*, 10:739, 2019.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. 11 2014.

[21] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. pages 2408–2415, 2012.

[22] Achraf Oussidi and Azeddine Elhassouny. Deep generative models: Survey. *2018 International Conference on Intelligent Systems and Computer Vision, ISCV 2018*, 2018-May:1–8, 5 2018.

[23] Taesung Park, Ming Yu Liu, Ting Chun Wang, and Jun Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:2332–2341, 3 2019.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2 2021.

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 10 2020.

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 4 2022.

[27] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *33rd International Conference on Machine Learning, ICML 2016*, 3:1681–1690, 5 2016.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 12 2021.

[29] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision 2000 40:2*, 40:99–121, 11 2000.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. 5 2022.

[31] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In *ICLR*, 2017.

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, Jenia Jitsev, and Uc Berkeley. Laion-5b: An open large-scale dataset for training next generation image-text models. *Research Center Juelich*, page 7, 6 2022.

[33] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

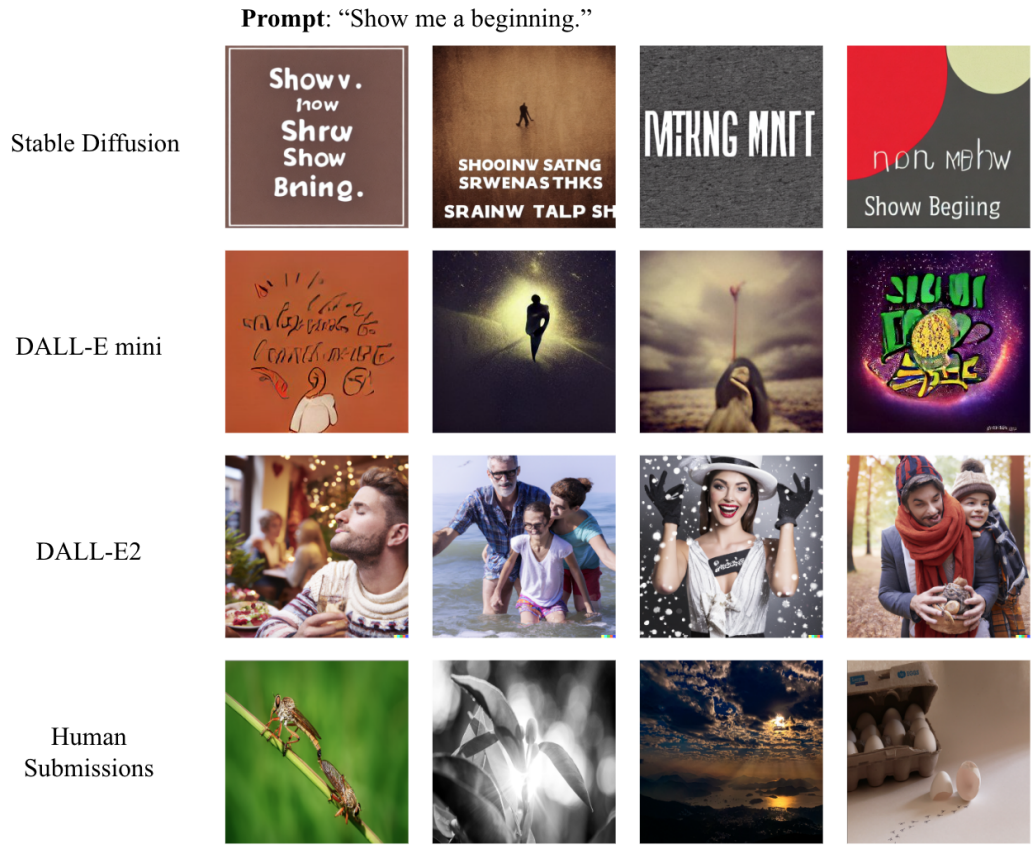

Figure 2: The prompt is deliberately open to interpretation, but these systems do not seem to have captured the underlying abstraction of the term "beginning".

[34] Leo Tolstoy. *What is Art?* Penguin, 1995[1897].

[35] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. 9 2016.

[36] Benjamin Walter. *The Work of Art in the Age of Mechanical Reproduction*. Penguin Books, 2008.

A Instances of prompts from AVA

Here we present some instances that we believe showcase one of the current strongest limitations of DGMs: their dependence on objective, fine-grained descriptions of the image in order to generate high-quality, semantically-similar images. To illustrate this limitation, we feed the model with prompts taken from the instructions given by the organisers of photographic contests at `https://www.dpchallenge.com` to participants, and compare the outcome of some popular diffusion models against human submissions. As it can be seen from Figure 2 and Figure 3, humans naturally understand figurative speech and creatively interpret the prompt in different ways. Diffusion models seem to struggle to go beyond the literal meaning.

B Wasserstein distance between datasets

We chose the Wasserstein distance to provide an intuition about how much it would require to convert one distribution into another, and compute it for each pair of datasets in the three variables considered

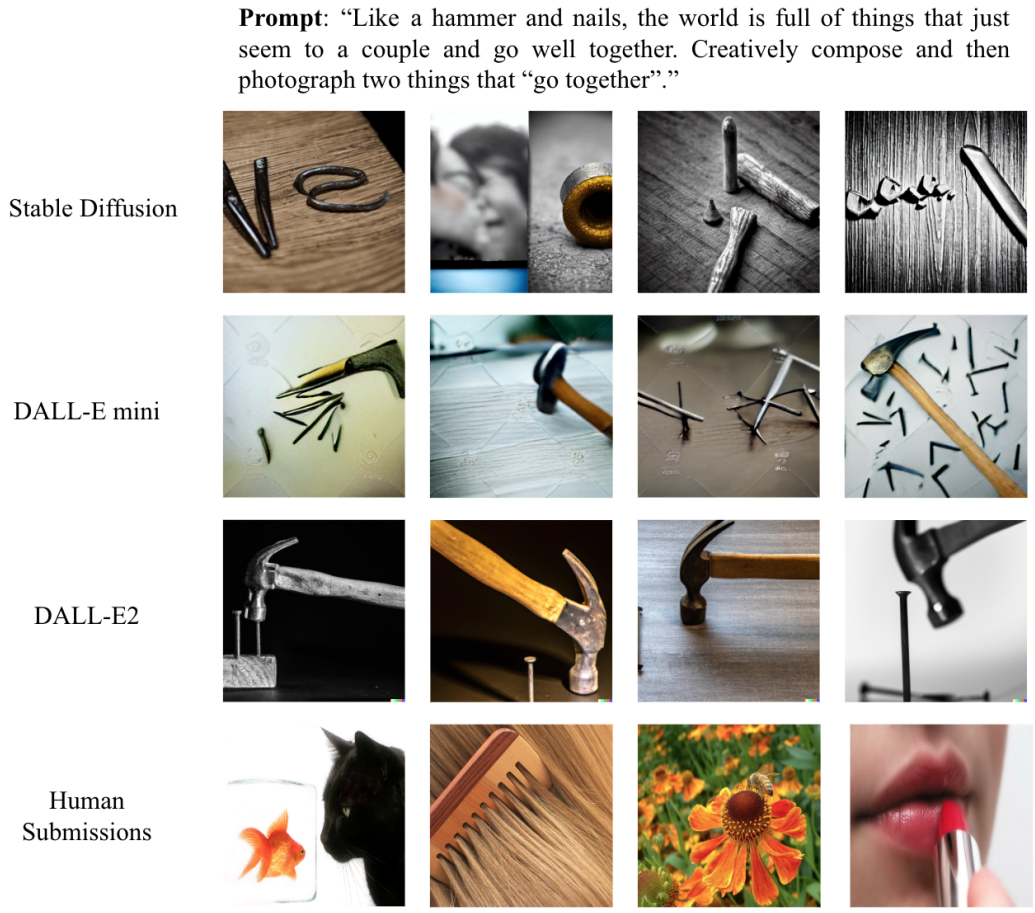

Figure 3: In this case the prompt suggests an example that participants can take figuratively to come up with new ideas, but instead diffusion models apparently understand only the literal meaning.

(sentimentality, subjectiveness and concreteness), as show in Table 1. The Wasserstein distance between two distributions u and v can be computed as:

$$W_1(u,v) = \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| d\pi(x,y) \tag{1}$$

where Γ is the set of (probability) distributions on $\mathbb{R} \times \mathbb{R}$ whose marginals are u and v on the first and second factors respectively.

We find that the average distance to shift from any distribution to another within datasets related to art is smaller ($\mu(W_1) = 2.71e^{-3}$, $\sigma^2(W_1) = 0.6e^{-3}$) than that required to transform it into either MS-COCO, Conceptual Captions, Captions12M or Laion-Aesthetics. An analogous effect is observed within these datasets ($\mu(W_1) = 3.16e^{-3}$, $\sigma^2(W_1) = 1.38e^{-3}$).

		Artemis	SemArt	MS-COCO	Conc. Caps.	Captions12M	Laion-Aesth.
Sentiment	AVA	3.53e-3	2.34e-3	9.51e-3	6.08e-3	4.24e-3	7.41e-3
	Artemis	-	2.25e-3	1.23e-2	9.15e-3	7.12e-3	1.04e-2
	SemArt		-	1.17e-2	8.09e-2	6.14e-3	9.4e-3
	MS-COCO			-	3.7e-3	5.76e-3	2.61e-3
	Conc. Caps.				-	2.23e-3	1.37e-3
	Captions12M					-	3.3e-3
Subjectivity	AVA	1.9e-3	4.55e-3	5.25e-3	5.49e-3	3.73e-3	6.98e-3
	Artemis	-	4.1e-3	6.29e-3	6.76e-3	4.77e-3	8.24e-3
	SemArt		-	8.83e-3	8.78e-3	6.79e-3	9.78e-3
	MS-COCO			-	1.17e-3	2.48e-3	2.43e-3
	Conc. Caps.				-	2.25e-3	1.55e-3
	Captions12M					-	3.47e-3
Concreteness	AVA	1.46e-3	1.46e-3	4.00e-3	3.69e-3	3.90e-3	5.8e-3
	Artemis	-	1.09e-3	3.69e-3	3.13e-3	3.35e-3	5.55e-3
	SemArt		-	2.76e-3	2.45e-3	2.62e-3	4.55e-3
	MS-COCO			-	8.37e-4	1.02e-3	1.99e-3
	Conc. Caps.				-	4.95e-4	2.43e-3
	Captions12M					-	2.27e-3

Table 1: Wasserstein distance (W_1) computed between each pair of datasets and language property considered in the study.