Measuring Diversity in Co-creative Image Generation

Francisco Ibarrola and **Kazjon Grace**
School of Architecture, Design and Planning
The University of Sydney
Sydney, Australia
[francisco.ibarrola, kazjon.grace]@sydney.edu.au

Abstract

Quality and diversity have been proposed as reasonable heuristics for assessing content generated by co-creative systems, but to date there has been little agreement around what constitutes the latter or how to measure it. Proposed approaches for assessing generative models in terms of diversity have limitations in that they compare the model's outputs to a ground truth that in the era of large pre-trained generative models might not be available, or entail an impractical number of computations. We propose an alternative approach based on entropy of neural network encodings for assessing diversity of sets of images that does not require ground-truth knowledge and is easy to compute. We also compare two pre-trained networks and show how the choice relates to the notion of diversity that we want to evaluate. We conclude with a discussion of the potential applications of these measures for ideation in interactive systems, model evaluation, and more broadly within computational creativity.

Introduction

The quality of generative AI text-to-image systems is improving rapidly, whether you assess that by fit-to-prompt, perceived realism, Frechét Inception Distance (FID) (Heusel et al. 2017), or by virtually any other measure. In a computational creativity context, however: is "quality" all we need? Especially in an interactive and/or co-creative context, that seems to be a dangerous assumption, given the long history in our field of creativity definitions featuring at least a duopoly of constituent factors: one broad cluster of factors that have been variously referred to as value, utility, appropriateness or quality, and another broad cluster usually referred to as novelty, originality, or surprise. The focus on quality is natural given the rapid advances of these technologies, but several specific questions arise when considering the sufficiency of image quality in our context: What happens when the prompt is ill-formed, because a user doesn't yet know what they want? What happens when a generator is asked to produce something that deviates substantially from the training data? What happens when a generator's understanding of a word or phrase differs from the user's?

Algorithmic measures of quality do not – at least at present – offer a way to address any of those questions. Conversely, using human subjects evaluation is too slow and too expensive to be a feasible source of feedback at the scale required to improve the underlying models. In this paper we propose that, at least for interactive contexts, generator quality (usually defined as some combination of matching the distribution of the data and accurately reflecting any conditioning stimuli such as prompts) must be accompanied by generator diversity (which we broadly define as maximising the breadth of options among the outputs, although we provide a more specific entropy-based definition below). The importance of generators offering users diverse options has been raised in computational creativity before, particularly in the field of procedural content for games (Smith and Whitehead 2010), we seek to expand those notions to cover all interactive generative AI, at least where output creativity is a potential goal. Armed with such a measure of generator diversity, we might begin more-systematically addressing those broader questions about generative AI and co-creative systems.

Diversity has been proposed (Preuss, Liapis, and Togelius 2014; Ibarrola, Lulham, and Grace 2023) as desirable in co-creative systems in the past, and in this paper we expand on those proposals, arguing for the criticality of diversity measures in interactive generative systems. We contend that the within-set diversity of generated content is a useful counterpart to quality in evaluating interactive text-to-image systems, formalise the problem of measuring it, and then present generalisable algorithms for doing so. By "within-set diversity" we refer abstractly to the breadth of a set of responses provided to a user as part of a single "round" of generation. Our motivation for this definition is that 1) creative tasks are by definition ill-specified and effective (human) approaches to combating that typically involve reframing/reformulation (Schön 1992; Dorst 2015; Grace and Maher 2016), a general finding that has been replicated specifically in the text-to-image literature in the form of iterative prompt "engineering" (Oppenlaender 2022), 2) a creative system is unlikely to know in advance the direction in which its user might want to reformulate the "problem", and 3) in interfaces where users are presented with multiple options to choose from (common in text-to-image UIs), traditional single-artefact models of novelty or surprise may result in duplicate content.

Our focus on diversity as a desirable quality for results produced during an on-going creative process is further motivated by research from the Information Retrieval (IR) community, which has long explored the utility of diversity as an accompaniment to accuracy/similarity in retrieving sets of search or recommendation system results (Candillier et al. 2011; Kunaver and Požrl 2017). In the IR community, the goal is to maximise the chances that an answer to the user's query exists within the top N results – for a concrete example, consider that "top N" to be on the "first page" of a search engine. To do so, it's not sufficient to present a set of most-similar or most-accurate results, because there's a high likelihood that those will be all self-similar: in other words, the within-set diversity of the results would be low. In the context of search and recommendations, this represents a poor use of the available screen real estate, for if one guess is wrong, all results are useless. We propose that this finding generalises to interactive text-to-image systems, and furthermore suggest that maximising within-set diversity alongside whatever measure(s) of quality are useful in context should – in theory – increase the potential for problem reframing and/or transformationally creative output.

It's important to note that our approach does not assume that there is a "ground truth" for image-set diversity: how people perceive the differences between objects is clearly subjective, constrictive, and situational. There is a vast number of ways that any two things could be compared, and the number of comparisons within or sets of things can only be greater. Our goal is not to chase an imaginary objective measure of perceived diversity, but instead to explore approximate measures that are sufficiently accurate at the population level to guide future research. This paper explores definitions for such a measure, concluding with our plans to empirically validate it.

With those assumptions in mind, it becomes necessary to precisely operationalise within-set diversity for an interactive text-to-image context. The field of quality-diversity algorithms (Pugh, Soros, and Stanley 2016), a form of multi-objective optimisation which has been extensively applied before in computational creativity (Zammit, Liapis, and Yannakakis 2022; Mccormack, Cruz Gambardella, and Krol 2023; Demke et al. 2023), would seem to be somewhere to look for inspiration, yet in those cases "diversity" is typically measured along several domain-specific and pre-defined behavioural variables: they offer no general measure of diversity that might be applicable in our context. Some generalisable ways of measuring diversity have already been proposed in the literature (Naeem et al. 2020), yet most of them either require a ground truth (a.k.a. access to a specific test dataset) for comparison, and/or are very expensive to compute. In today's era of large pre-trained generative models, access to a dataset representative of the generator's target distribution cannot be assumed, and there is a need for a scalable, general, dataset-blind measure of within-set diversity.

To overcome these issues, we propose and compare two versions of a more relaxed approach to estimating within-set diversity that can be computed quickly and without knowing the distribution of the training data. Our approach is instead based on general pre-trained network mappings. Having no ground truth to evaluate our own measures, we also propose an approach for generating artificial data which we would expect a-priori to exhibit a pattern of relative diversity levels, allowing us to check whether the proposed methods align with our expectations. We argue that our proposed measures are more useful than the state-of-the-art in terms of practicality, particularly in the domain of high-quality interactive image generation in computationally creative contexts.

Methods

When deep neural networks are trained for image classification or similar tasks, the data from an image flows from each layer to the next as a tensor of values usually referred to as layer "activations". These activations contain information about the different characteristics of each image, and are in turn interpreted by the following layer. Since activations are learnt to be useful for performing the task for which the network was trained – general purpose image recognition, generation, or segmentation, for example – then pre-trained networks are often "cropped" at certain layers, allowing those layer activations to be used as the input to train (typically smaller) networks for different purposes (Kora et al. 2022).

This idea has been exploited for other uses, such assessing the quality of image generators using FID (Heusel et al. 2017). This method uses the second-to-last layer activation of the general-purpose pre-trained image network InceptionV3 (Szegedy et al. 2016) as latent variables, effectively casting them as constituting a "conceptual space" of all natural images (Boden 2004). FID then compares a test set of real images to a set of generated ones, with a "perfect" score of 0 indicating that the distribution of features in the generated images is identical to those of the "real" ones. Specifically, under a normality hypothesis, the Frechét Distance between the empirical distributions of the latents can be computed explicitly, giving a good proxy for the quality of the generative process.

While this provides a reliable assessment of the ability of a generator to match a dataset, it has two drawbacks. Firstly, that diversity cannot be measured directly (and in fact moving away from the original latent distribution by becoming "more diverse" will produce worse FID scores). And secondly, this method requires a ground-truth distribution for computing (a.k.a. a dataset of all relevant "real" images), which as previously discussed is inconvenient for our purpose.

Nonetheless, the idea of analyzing an image dataset through the latent space of a pre-trained network can still be of use. By analyzing the (empirical) probability distribution of a set of generated images, we may get an idea of diversity by looking at its entropy, which has been widely used as a diversity index (Jost 2006) in other fields. Where FID computes quality as the distance between the distributions of a generated set and a ground truth in a latent space, we instead seek to assess diversity as the entropy of the generated set's same latent variables.

We describe two approaches to doing so below, detailing how to tractably approximate entropy in a co-creative case. Both measures are "truncated" in that they use approximate measures of entropy in order to avoid the requirement of having at least as many samples (as in generated images) as the dimensionality of the latent space, which isn't feasible in most interactive contexts. The first, Truncated Inception Entropy (Ibarrola, Lawton, and Grace 2022), is a measure of diversity using the same latent space as in the broadly-adopted FID, on the motivation that if the second-to-last layer of the Inceptionv3 model is a good proxy for image features relevant to quality, it should likely be similar with respect to diversity. In the second, Truncated CLIP Entropy, we instead explore the use of Contrastive Language-Image Pre-Training (CLIP (Radford et al. 2021)), a multi-modal embedding of both text and images. This is motivated by the assumption that in some use-cases, diversity in a "semantic" space that can embed both prompt and images may be more relevant than the features of a general-purpose image model.

Truncated Inception Entropy

Let us consider a function f that maps images into a latent space $\mathcal{Z} \subset \mathbb{R}^D$, in such a way that the points have a normal distribution $\pi_f \sim \mathcal{N}(\mu, \Sigma_i)$ in $\mathcal{Z}$. This normality assumption is the same one used when computing FIDs, where f is a truncated version of the InceptionV3 network on the last layer, with output size $D = 2048$.[1] The resulting latent space with this choice of network is thus $\mathcal{Z} = \mathbb{R}^{2048}$.

Under this hypothesis, we could assess the diversity of a given set of images A by feeding them to the truncated InceptionV3 network to get their corresponding (normally distributed) latents, and then computing the differential entropy h (Shannon 1948), defined as

$$h(\pi_f) = -\mathbb{E}\log(\pi_f) = \frac{1}{2}\log\det(2\pi e\Sigma_i). \tag{1}$$

When the number of samples N in a set of images A is smaller than the dimension D of the latent space, the empirical approximation $\hat{\Sigma}_i$ of Σ_i is singular, meaning that the determinant is null and hence the latter computation unfeasible. To overcome this, it has been proposed (Ibarrola, Lawton, and Grace 2022) that a truncated version of entropy can be used, defined as

$$\mathrm{TIE}_K(A) \doteq \frac{K}{2}\log(2\pi e) + \frac{1}{2}\sum_{k=1}^{K}\log\lambda_k^{(i)}, \tag{2}$$

where TIE denotes Truncated Inception Entropy, and $\{\lambda_k^{(i)}, k = 1, \ldots, K\}$ is the set of the K largest eigenvalues of $\hat{\Sigma}$. Note that $K = D$ would make the TIE equivalent to Equation (1), but choosing a smaller value for K would let us compare diversities of smaller sets of images.

Truncated CLIP Entropy

The InceptionV3 network was trained as a classifier over the ImageNet database (Deng et al. 2009). Since then, new pre-trained networks have been made available, such as CLIP, in which images and text are encoded together in a shared latent space $\mathcal{Z} \subset \mathbb{R}^{512}$. This is done in such a way that text or images with the same semantic characteristics are grouped together, which may be a useful feature of a space in which we want to calculate within-set diversity.

In an analogous way as with the TIE, we may consider a set A of N images and $g(A) \doteq \{g(a), a \in A, g(a) \in \mathbb{R}^{512}\}$ set of latent CLIP representations of the images (where g denotes the CLIP image encoder). From this set, we can calculate the empirical covariance matrix $\hat{\Sigma}_c \in \mathbb{R}^{512\times512}$, and subsequently its K largest eigenvalues $\{\lambda_k^{(c)}, k = 1, \ldots, K\}$ to compute the Truncated CLIP Entropy (TCE) as

$$\mathrm{TCE}_K(A) \doteq \frac{K}{2}\log(2\pi e) + \frac{1}{2}\sum_{k=1}^{K}\log\lambda_k^{(c)}. \tag{3}$$

Note that while the computation is the same as that of the TIE, the values are not directly comparable, since the spaces in which the InceptionV3 and CLIP latents are defined are different (hence the supra-index notation on the eigenvalues).

Open-source Implementation

The code (Python3) for trying the measures described here is freely available, and may be installed using pip

```
$ pip install image-diversity
```

and tested by running

```
$ python3 image_diversity <path/to/dir>
```

where `<path/to/dir>` is a path to a directory containing a set of images to be evaluated.

More details on installation and usage can be found at https://github.com/fibarrola/image_diversity

Experiments

Comparing diversity as estimated by either TIE or TCE is a non-trivial problem, given that there is no ground truth on what the diversity of set of images "should be". Or rather, what makes a set of images more or less diverse than another. We are currently in the process of designing a set of human-subjects evaluations to compare different versions of these measures on the degree to which they align with human evaluation. A key challenge in that experimental design is what exactly to ask people to do, rate, or judge in order to validate our diversity measures and our hypothesis that generator diversity facilitates output creativity. For this paper, however, we present a series of in-silico experiments. We have built sets of images using different processes that we judge should lead to more or less diversity, and confirmed whether our diversity measures reflect those a-priori assumptions. This approach is consistent with past experiments on computational diversity measures, such as in the domain of text documents (Bache, Newman, and Smyth 2013). Specifically, we automated the generation of sets of prompts that vary in content and style in ways that are both congruous and incongruous.

This was carried out using GPT-3.5 (Brown et al. 2020) to generate different text prompts, which were in turn used to generate five datasets: Control with Low Noise (a fixed

[1] This choice of layer from which to extract activations is the standard for FID, yet other intermediate layers might be considered provided a reliable way to deal with their high dimensionality. Further exploration is required.

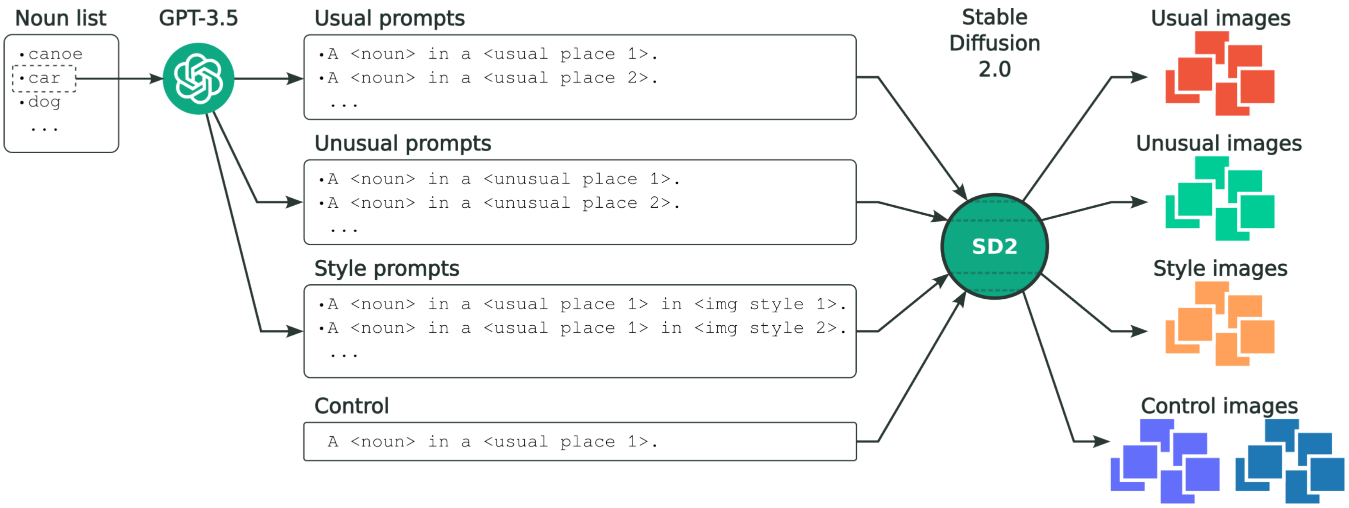

Noun list
•canoe
•car
•dog
...
GPT-3.5
Usual prompts
•A <noun> in a <usual place 1>.
•A <noun> in a <usual place 2>.
...
Unusual prompts
•A <noun> in a <unusual place 1>.
•A <noun> in a <unusual place 2>.
...
Style prompts
•A <noun> in a <usual place 1> in <img style 1>.
•A <noun> in a <usual place 1> in <img style 2>.
...
Control
A <noun> in a <usual place 1>.
Stable Diffusion 2.0
SD2
Usual images
Unusual images
Style images
Control images

Figure 1: Image set generation process for diversity evaluation.

prompt with small variations in random generative components), Control with High Noise (a fixed prompt with large variations in random generative components), Usual (a given object in different places it might be), Unusual (a given object in places it would not be) and Style (a given object in a Usual place rendered in different visual styles).

If our intuitions about within-set diversity are accurate, two things should occur. Firstly, the low noise Control set should be less diverse than the high noise Control set. Secondly, the low noise Control set should show less diversity than the Usual set, and both of them less than the Unusual set. Finally, the Style set's diversity should be purely visual with low semantic variations, and hence we expect it might be assessed differently by TCE and TIE, due to the latter's presumed greater reliance on visual rather than semantic differences.

The generative process of the image sets is illustrated in Figure 1, and was conducted as follows. We first chose five nouns: *canoe*, *car*, *dog*, *coffee mug* and *pigeon* and then gave the LLM instructions to generate three different sets of prompts as follows:

Usual: Generate a list of 45 places where a [noun] may be. Print as "A [noun] in <place>"

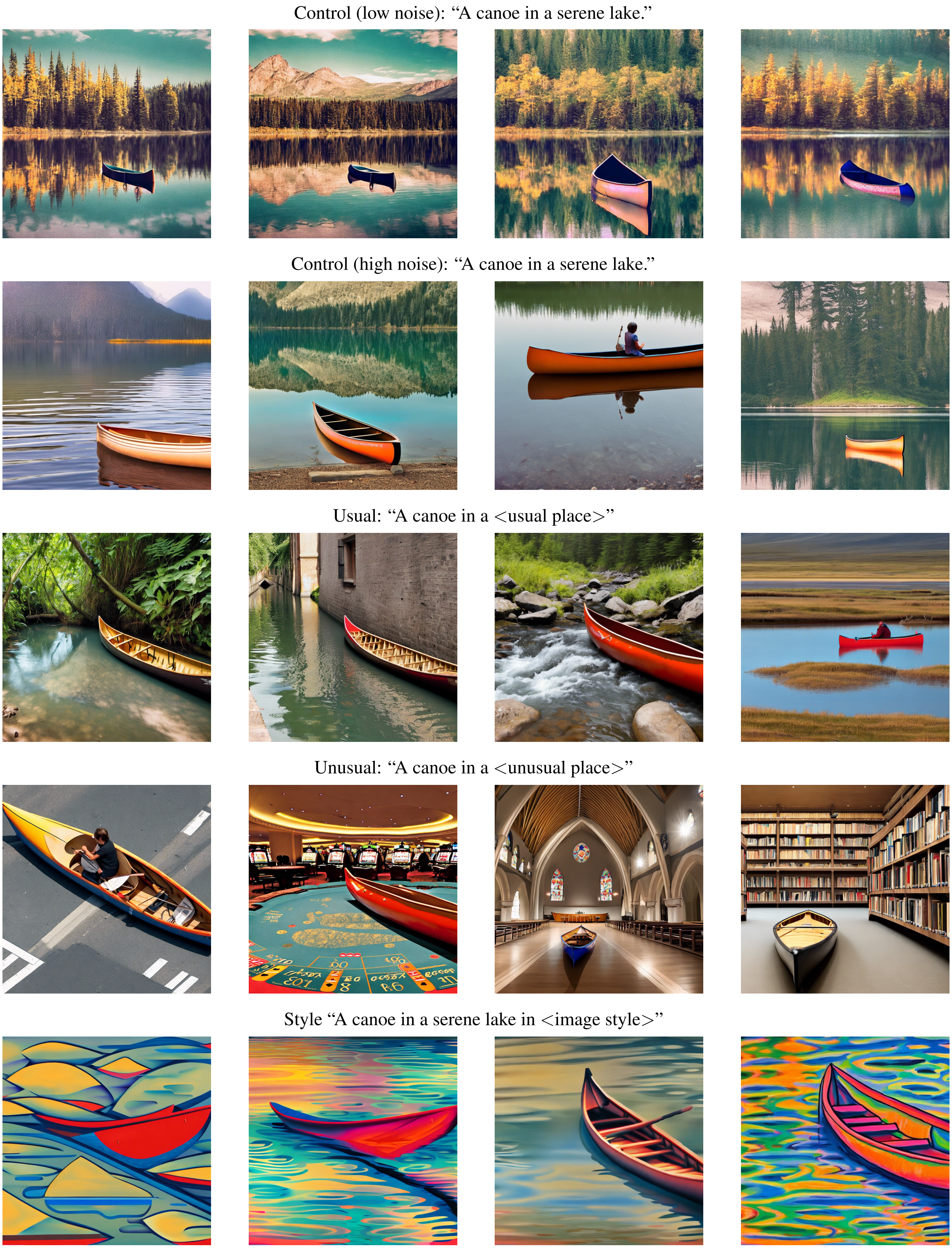
Unusual: Generate a list of 45 places where finding a [noun] would be absurd. Print as "A [noun] in <place>"

Style: Generate a list of 45 painting or image styles. Print as "A [noun] in [place] in <style>"

In each case [noun] was replaced with one of the five objects in the list above. The *Control* sets did not use an LLM to generate as the prompts were fixed to a single place, chosen to be stereotypical for that object. The prompts in this case were *a canoe in a serene lake*, *a car in a driveway*, *a dog in a backyard*, *a coffee mug in an office*, and *a pigeon in a tree*. These same "fixed" places were used for each of the Style prompts. All the images of these three sets were generated using the same random parameters (or seed) so that the prompts are the only source of variability. In the Control sets, since the prompts were fixed, the variations were obtained by letting the initial random noise parameters to change (with fixed noise all the images would have been identical). In the high-noise set these were completely random, while in the low-noise Control set, the parameters were built around a random mean with 20% variance, resulting in random values much close to each other across the set than those on the high-noise Control group. The effect of this can be observed on the samples in Figure 2.

Finally, we used Stable Diffusion (Rombach et al. 2022) to generate sets of 45 images per prompt, examples of which can be seen in Figure 2. It can readily be observed that the low-noise Control set show images very similar both visually and in terms of the depicted elements, while the high-noise Control set varies much more visually, yet no more in terms of elements. The Usual set shows some common elements besides the canoe itself, such as water and vegetation, but more variability than the control. In contrast, the Unusual set shows a variety of elements not quite related to each other, from which we should expect a larger diversity. Finally, the Style set is quite consistent in terms of the depicted scene, but is more diverse in terms of geometry and textures.

For each of the 5 objects, we built 10 random subsets of 30 (out of 45) images, and computed the TIE and TCE values, depicted in Figure 3. It can be seen that, as expected, the low-noise Control set shows lower diversity than the rest, and the highest diversity scores are observed for the Unusual set with both methods. Unsurprisingly, also, reducing the variance of the input noise (in the Control set) reduces the diversity of the output. However, the TIE marks the Unusual and Style sets as having comparable diversity, significantly greater than the mild variations in the Usual and Noise groups, whereas the TCE tells a different story. In this case, the variations in visual style carry a lower weight than those of the elements composing the image, meaning



Control (low noise): “A canoe in a serene lake.”

Control (high noise): “A canoe in a serene lake.”

Usual: “A canoe in a <usual place>”

Unusual: “A canoe in a <unusual place>”

Style “A canoe in a serene lake in <image style>”

Figure 2: Samples of image sets generated with one of three methods to evaluate diversity behaviour.

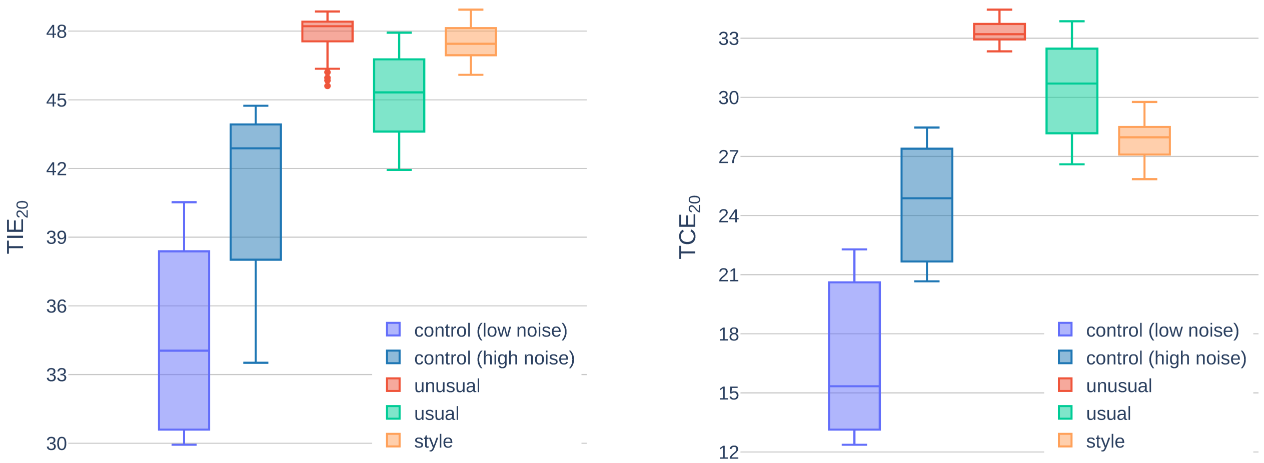


Figure 3: Diversity values (TIE and TCE) using $K = 20$ eigenvalues, for sets of images generated with four different criteria. All the between-set differences are statistically significant $(p < 0.01)$ except for the TIE for the *unusual* and *style* sets. Note that the obtained diversity values are not comparable between methods on account of using different latent spaces.

that TIE and TCE are accounting for diversity in two different senses. This comports with our expectation that TCE "weights" semantics higher in its accounting of image diversity.

Text diversity

As mentioned before, the CLIP network on which TCE is based embeds both images and text in a shared latent space. This means that TCE can be computed (as in 3) on the CLIP latents of a set of prompts directly, without requiring that they be first converted into images. This suggests a potential application of TCE to text diversity, which may be useful by itself or as a comparison to image diversity.

While a rigorous evaluation would be required before claiming that TCE could be used on text to assess semantic diversity in any useful way, we conducted a preliminary experiment of computing the TCE over the prompts (see Figure 1) used in our previous experiments, with the exclusion of the Control sets for which the prompts were all identical.

The results are depicted in Figure 4 and are broadly in line with those obtained for images for the Usual and Unusual groups, with the latter being higher. It can also be observed that there is a very wide gap between these and the Style set, which was not observed in the case of images. This makes sense, as the prompt texts only differed by that one or two style words, making them semantically quite similar, while that one word had a large effect on the visual content of the image, at least according to TIE. This again provides some early evidence to support our diversity measures as capturing a quantity of potential interest to the developers of co-creative systems and other interactive applications of generative AI.

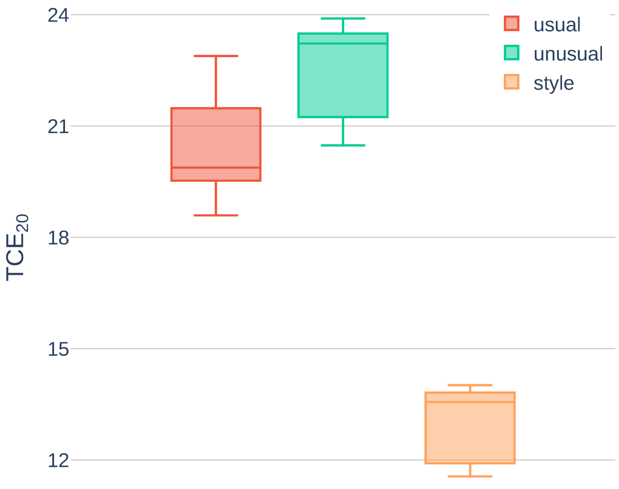


Figure 4: TCE using $K = 20$ eigenvalues, for sets of text prompt generated with three different criteria. All the between-set differences are statistically significant $(p < 0.01)$.

Discussion and Conclusions

We proposed a method to assess diversity in image datasets that is agnostic to training data and simple to compute. The method was compared to its analogous using another network's latent space, and results show both variants to align well with expected outcomes. Furthermore, it has been shown that the different networks assess diversity in different senses, meaning that they might serve for different creative contexts.

Given that we have not yet validated that our method correlates with the subjective perceptions of human subjects, it's difficult to conclusively say it might help the design of future co-creative systems. While that validation is our next step, our broader hypothesis is that generators with greater diversity in their responses to typical user requests would be more effective at augmenting creativity. For example, an AI sketching system might show a user a set of very different completions of their partial drawings. Or a text-to-image system might begin a session by generating highly diverse responses to aid in iterating on the prompt. Or an AI story-writing system might suggest wildly different branching futures for a partially complete story as a way to combat writers' block. In each of these examples the creative act is still in the early "divergent thinking" stages, where to focus on producing a high-quality solution might mean to converge prematurely. In the extreme case, improvements in generator diversity may even be of greater utility than further improvements in generator accuracy, although that claim would also need to be validated.

Our measures are based on approximations of entropy, and entropic measures of diversity have faced some criticism in other fields, such as in biology (Jost 2006). The criticism is that the actual quantity of interest in diversity is how many meaningfully active categories (species in biodiversity, "features" in an image) in a sample, not the amount of information required to identify which category a randomly-selected sample belongs to. Qualities such as *balance*, *variety* and *disparity* have been proposed as necessary components of this kind of categorical measure of diversity (Stirling 2007). This approach has been applied to evaluating document diversity using topic modelling to generate the categorical representation (Bache, Newman, and Smyth 2013). In the case of image generation, this might suggest an alternative formulation in terms of the number of features identified by some appropriately categorical representation.

While our results are promising, further experiments are needed to fully assess the proposed methods' compliance with expectations in creative computing applications. Particularly, future work shall deal with the validation of these metrics in comparison with human perception, and exploring the use of latent spaces of other pre-trained neural networks. In fact, the possibility of using average pooling for computing FID using intermediate InceptionV3 layers has been proposed, although not properly tested (Seitzer 2020), and its usage for computing TIE is thus equally plausible. Using earlier layers in the image encoding network as the latent space in which diversity is calculated could yield a more texturally- or visually- biased measure, which may be useful for some scenarios, although only if some technique like average pooling can be applied to reduce their dimensionality.

It's also important to consider the potential limitation of using an LLM – in our case GPT3.5 – to generate sets of presumably-diverse prompts. It's likely that these sets of prompts are biased in ways that are hard to quantify, potentially harming the generalisability of our conclusions. However, from manually inspecting the lists of prompts, we can say that GPT3.5 seems less biased than we the authors would be if asked to manually construct a list of 45 places where an object (e.g. a canoe) should or shouldn't be. Additionally, the experiments described in this paper do not require the prompts to be an unbiased sample of language, since we are comparing sets produced by the same generator. Nevertheless, the suitability of this approach should be considered as we go forward with human evaluations.

Finally, as shown by the preliminary experiments, it is worth noting that TCE might also be used to assess text diversity on account of the CLIP latent space being the same for either text or images. More experiments are needed to properly test whether or not this works reliably in practice, contrasting it with other text diversity assessment methods. Our current research is exploring both the design of those experiments as well as the design of future generative systems aimed at producing small sets of diverse-yet-high-quality responses for use in co-creative systems.

Acknowledgments

We would like to acknowledge the Australian Research Council for funding this research (ID #DP200101059).

References

[Bache, Newman, and Smyth 2013] Bache, K.; Newman, D.; and Smyth, P. 2013. Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 23–31.

[Boden 2004] Boden, M. A. 2004. *The creative mind: Myths and mechanisms.* Psychology Press.

[Brown et al. 2020] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

[Candillier et al. 2011] Candillier, L.; Chevalier, M.; Dudognon, D.; and Mothe, J. 2011. Diversity in recommender systems. In *Proceedings: The Fourth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services. CENTRIC*, 23–29.

[Demke et al. 2023] Demke, J.; Grace, K.; Ibarrola, F.; and Ventura, D. 2023. Transformational creativity through the lens of quality-diversity. In *Proceedings of ICCC'23*.

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

[Dorst 2015] Dorst, K. 2015. *Frame innovation: Create new thinking by design*. MIT press.

[Grace and Maher 2016] Grace, K., and Maher, M. L. 2016. Surprise-triggered reformulation of design goals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

[Heusel et al. 2017] 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.

[Ibarrola, Lawton, and Grace 2022] Ibarrola, F.; Lawton, T.; and Grace, K. 2022. A collaborative, interactive and context-aware drawing agent for co-creative design. *arXiv preprint arXiv:2209.12588*.

[Ibarrola, Lulham, and Grace 2023] Ibarrola, F.; Lulham, R.; and Grace, K. 2023. Affect-conditioned image generation. *arXiv preprint arXiv:2302.09742*.

[Jost 2006] Jost, L. 2006. Entropy and diversity. *Oikos* 113(2):363–375.

[Kora et al. 2022] Kora, P.; Ooi, C. P.; Faust, O.; Raghavendra, U.; Gudigar, A.; Chan, W. Y.; Meenakshi, K.; Swaraja, K.; Plawiak, P.; and Acharya, U. R. 2022. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering* 42(1):79–107.

[Kunaver and Požrl 2017] Kunaver, M., and Požrl, T. 2017. Diversity in recommender systems–a survey. *Knowledge-based systems* 123:154–162.

[Mccormack, Cruz Gambardella, and Krol 2023] Mccormack, J.; Cruz Gambardella, C.; and Krol, S. 2023. Creative discovery using quality-diversity search. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, 747–750.

[Naeem et al. 2020] Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 7176–7185. PMLR.

[Oppenlaender 2022] Oppenlaender, J. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, 192–202.

[Preuss, Liapis, and Togelius 2014] Preuss, M.; Liapis, A.; and Togelius, J. 2014. Searching for good and diverse game levels. In *2014 IEEE Conference on Computational Intelligence and Games*, 1–8. IEEE.

[Pugh, Soros, and Stanley 2016] Pugh, J. K.; Soros, L. B.; and Stanley, K. O. 2016. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI* 3:40.

[Radford et al. 2021] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

[Rombach et al. 2022] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

[Schön 1992] Schön, D. A. 1992. Designing as reflective conversation with the materials of a design situation. *Knowledge-based systems* 5(1):3–14.

[Seitzer 2020] Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`. Version 0.3.0.

[Shannon 1948] Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal* 27(3):379–423.

[Smith and Whitehead 2010] Smith, G., and Whitehead, J. 2010. Analyzing the expressive range of a level generator. In *Proceedings of the 2010 workshop on procedural content generation in games*, 1–7.

[Stirling 2007] Stirling, A. 2007. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society interface* 4(15):707–719.

[Szegedy et al. 2016] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

[Zammit, Liapis, and Yannakakis 2022] Zammit, M.; Liapis, A.; and Yannakakis, G. N. 2022. Seeding diversity into ai art. *arXiv preprint arXiv:2205.00804*.