ADCA: Artifact-Based Dataset Creativity Assessment

Harrison Sims
University of Michigan
Ann Arbor, MI 48109
harsims@umich.edu

Gabriel Ganberg, Robert McCormack, Svitlana Volkova
Aptima Inc.
8 Cabot Road, Suite 4000, Woburn, MA 01801
gganberg,rmccormack,svolkova@aptima.com

Abstract

We present a three-dimensional framework for automated dataset creativity assessment that decomposes creativity into measurable components: attribute novelty, recombination novelty, and feature addition. Our method treats data points as collections of categorized artifacts, evaluating universal and unique attributes through semantic embedding comparison. Attribute novelty measures semantic diversity using pairwise cosine similarity of CLIP embeddings, recombination novelty quantifies unique attribute co-occurrences via hierarchical clustering, and feature addition assesses unique embellishment distribution. Validation on three 100-image datasets shows significant statistical differences across metrics ($p < .001$), with forced creativity datasets achieving the highest attribute novelty and prompted creativity datasets demonstrating superior recombination patterns. Strong positive correlations between metrics ($r = 0.43 - 0.55$) support construct validity. This modality-agnostic, embedding-based evaluation framework enables systematic assessment of generated image data quality beyond traditional performance benchmarks, with direct implications for foundation model training in high-stakes applications. The three 100-image datasets are publicly available[1].

1 Introduction

Operational deployment of computer vision systems for ground vehicle recognition and overhead imagery analysis [1–3] requires training datasets with sufficient diversity/creativity to ensure robust performance across varied conditions [4–6]. Current dataset evaluation methodologies rely on single-metric assessments that fail to capture the multifaceted nature of visual creativity required for reliable model performance. This limitation becomes critical when synthetic data generation augments limited real-world imagery collections [7], where inadequate creativity assessment results in models that perform well on benchmarks [8, 9] but fail during deployment [10–12].

Existing mathematical evaluation approaches [13, 14] predominantly use comprehensive single-score methodologies that aggregate diversity into a singular metric, which makes it difficult to isolate specific diversity characteristics that influence model robustness. Recent decomposition-oriented work in creativity evaluation [15, 16] has advanced analysis for generative models but remains oriented toward generator assessment rather than dataset evaluation. As a result, current methods do not fully address operational computer vision requirements in which training data must capture realistic environmental attribute co-occurrence patterns and contextual detail distributions.

The operational implications are acute in ground vehicle recognition systems, where models must reliably identify targets across diverse terrain, weather, and seasonal conditions. Overhead imagery analysis requires models trained on datasets capturing infrastructure configurations, vegetation patterns, and atmospheric conditions present in operational imagery. Current frameworks provide

[1] Datasets and evaluation pipeline are available at https://github.com/harrisonsims/ADCA

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

insufficient granularity to identify diversity gaps that compromise performance in high-stakes applications.

We introduce a novel three-dimensional framework for automated dataset creativity assessment through artifact-based decomposition. Our approach treats data points as collections of categorized information artifacts, evaluating universal attributes essential for task performance and unique attributes providing contextual detail. The framework decomposes creativity into attribute novelty (semantic diversity using embedding similarity), recombination novelty (unique attribute co-occurrences through hierarchical clustering), and feature addition (proportion and distribution of unique embellishments).

Key contributions include: (1) a modality-agnostic framework enabling creativity assessment for ground-based and overhead imagery datasets, (2) decomposition into interpretable components relating to operational performance requirements, (3) validation demonstrating statistical differentiation between datasets, and (4) correlation analysis supporting construct validity. This framework enables practitioners to systematically assess training data quality for mission-critical computer vision applications beyond traditional performance benchmarks.

2 Related work

With the growing popularity of LLMs and other artificial intelligence content generators, there has been considerable research in the space of autonomous evaluation. Most autonomous evaluation has been focused on generated text content, including the development of relevant question answering benchmarks like MMLU [17] and HellaSwag [18], and evaluation tools like BERTScore [19].

There have been parallel efforts in the image generation field focusing primarily on image-caption alignment, where the quality of autonomously generated captions, generated by VLMs like BLIP-2, is evaluated against ground-truth images. One of the most common types of text-to-image evaluation is embedding-based alignment evaluation [20], where the semantic meanings of texts and images are represented as numerical vectors that can then be directly compared using cosine distance. Despite being heavily embedder-specific, this method of evaluation is very popular because it is modality agnostic. Any two pieces of data, whether text, image, or otherwise, can be directly compared via their embeddings as long as there is an embedder specifically trained for the task. CLIPScore [21] is one of the most popular evaluation methodologies, where the semantic alignment between an image and its corresponding text caption is quantified according to the cosine similarity of the two CLIP embeddings.

The advancement of image evaluation tools like CLIP enabled the development of methodologies that measure the creativity of image generators. While not directly the same as dataset diversity, generator creativity uses many of the same fundamental theories and methods. There are two dominant taxonomies of creativity evaluation: subjective evaluation and mathematical evaluation. Subjective evaluation is when a judge (either human or a VLM) provides a subjective creativity score to generated content (typically a single image or a set of images). Human creativity assessment [22, 23] is considered the gold standard, but it is incredibly time-intensive and costly. In contrast, using a VLM trained on human-annotated datasets as a judge [24] may not be as widely accepted, but the automated nature allows for much better scaling as evaluation quantity increases. While the subjective evaluation taxonomy is more readily translatable to the human evaluation experience (which is inherently subjective), mathematical evaluation techniques allow for a much more transparent evaluation procedure. There are two primary mathematical approaches to the measurement of creativity: the comprehensive approach and the decomposition approach. The comprehensive approach uses a single score and method to measure creativity as a whole. This can involve running the generator through a set of tests and scoring the outputs [13] or using a single measure line semantic similarity to comprehensively operationalize creativity [14]. Despite the ease of calculation, the comprehensive approach risks oversimplifying creativity. In contrast, the decomposition approach breaks creativity down into several dimensions, like value, novelty, surprise, and cohesion, and scores each dimension [15, 16]. Despite the differences between the decomposition and comprehensive approaches, most mathematical methodologies leverage the same fundamental principle: artifact analysis. In many of these evaluation approaches, the image is broken down into parts called artifacts, each artifact is evaluated for semantic meaning using an embedder like CLIP [25], and the semantic meanings are compared to extract a difference or similarity score. The goal of the following study is to leverage some of the text-to-image evaluation tools and the decomposition approach to mathematical creativity

evaluation to develop a novel multi-dimensional dataset creativity evaluation methodology based on artifact analysis.

3 Evaluation metric definition

In 2010, Maher proposed the following dimensions for evaluating the creativity of a generator while in the process of generation: value, novelty, and surprise [26]. In 2025, Ramaswamy and colleagues expanded on Maher's work by proposing their own set of generator creativity dimensions: prompt requirement satisfaction, cohesion, and diversity [16]. When evaluating generated dataset creativity instead of generator creativity, there are key differences that need to be accounted for. Dimensions representing quality should be discarded, because there is a typically a value threshold for inclusion in the dataset. Furthermore, surprise becomes no longer relevant as expectation cannot be measured for a dataset where all the data is presented at once. This leaves only the dimensions of novelty, which should be specified to capture the different facets of dataset creativity. In our evaluation approach, we focus on three key dimensions of novelty: attribute novelty, which measures differences between individual artifacts; recombination novelty, which accounts for differences in co-occurring artifacts; and addition novelty, which accounts for the addition of unique artifacts.

Task formulation Evaluation dimensions are highly dependent on the downstream task. The two key task parameters are two sets of artifacts that define the downstream task: universal attributes and unique attributes. The set of universal attributes, $\mathcal{U}$, is a set of categories for artifacts that should appear in every generated image and be essential for image understanding. In contrast, the set of unique attributes, $\mathcal{Q}$, is a set of categories for non-essential artifacts that add detail and are expected to only appear in some images.

Consider an image dataset $\mathcal{D} = \{I_1, I_2, \ldots, I_N\}$ where N is the number of images in the dataset. For each image I_i in the dataset, the artifact extractor extracts the artifact instance set $\{a_{i,1}, \ldots, a_{i,L_i}\}$ where L_i is the number of artifacts identified in image I_i. Each artifact instance $a_{i,l} = (c_{i,l}, p_{i,l}, e_{i,l})$ is a tuple where: $c_{i,l} \in C = \mathcal{U} \cup \mathcal{Q}$ is the predicted category of the artifact instance, $p_{i,l} \in [0, 1]$ is the confidence score associated with category assignment, and $e_{i,l} \in [0, 1]^{512}$ is the semantic embedding of the artifact instance.

Attribute novelty Attribute novelty is the measure of the difference between artifacts of the same category C within a single dataset $\mathcal{D}$, and it only accounts for universal attributes. Consider J sets of embeddings E_j^A, where each embedding corresponds to an artifact categorized as universal attribute j and $n_j^A = |E_j^A|$. Attribute-wise attribute novelty score A_j is calculated by taking the average pairwise cosine similarity score of all the embeddings in set E_j^A following Eq. 1. Total attribute novelty score A is calculated by taking the average of the attribute-wise attribute novelty scores:

$$A_j = \frac{2}{n_j^A(n_j^A - 1)} \sum_{1 \le x < y \le n_j^A} \frac{E_{j,x}^A \cdot E_{j,y}^A}{\|E_{j,x}^A\| \, \|E_{j,y}^A\|} \tag{1}$$

Recombination novelty Recombination novelty is the measure of the frequency with which different pairs of attributes from different attribute categories occur in a single dataset $\mathcal{D}$. Similar to attribute novelty, recombination novelty only accounts for unique attributes. Consider the same J sets of embeddings, now notated as E_j^R where $n_j^R = |E_j^R|$. Each set of embeddings is clustered using a hierarchical method where the optimal set of clusters is returned.

Following the clustering of the J sets, each artifact $a_{i,l}$ where $c_{i,l} \in \mathcal{U}$ has a newly associated group classification $g_{i,l}$ based on the clustering. For each pair of universal attributes $(u_s, u_t \in \mathcal{U}, s \neq t)$ all of the appropriate pairs of artifact groupings $(g_{i,s'}, g_{i,t'})$ where $c_{i,s'} = u_s$ and $c_{i,t'} = u_t$ are extracted from the images in the dataset. For each pair (u_s, u_t), the proportion of unique pairs is calculated, and the recombination novelty score is calculated according to Eq. 2.

$$R = \frac{2}{J(J-1)} \sum_{s \neq t \in \mathcal{U}} \frac{\sum_i \mathbf{1}\{(g_{i,s'}, g_{i,t'}) \text{ are unique}\}}{\sum_i \mathbf{1}\{(g_{i,s'}, g_{i,t'}) \text{ both exist}\}} \tag{2}$$

Feature addition Feature addition is a measure of both the proportion of added features that are unique (aka unique score) and the proportion of images with an added feature in a dataset $\mathcal{D}$ (aka

incidence score). Feature addition only accounts for unique attribute categories because the unique attributes are meant to be inconsequential to the interpretation of the image and added for decoration. Consider the K sets of embeddings where E_k^F is the set with size n_k^F. Similar to recombination novelty, the embeddings of each unique attribute category should be hierarchically clustered into a set of G_k clusters according to a strict distance threshold τ, where feature addition unique score is calculated according to Eq. 3.

$$F_{\text{unique}} = \frac{\sum_{k=1}^{K} |G_k|}{\sum_{i=1}^{N} \sum_{l=1}^{L_i} \mathbf{1}\{c_{i,l} \in \mathcal{Q}\}} \tag{3}$$

The feature addition incidence score is simply the proportion of images with a unique feature and can be calculated following Eq. 4.

$$F_{\text{incidence}} = \frac{\sum_{i=1}^{N} \mathbf{1}\{\exists\, l : c_{i,l} \in \mathcal{Q}\}}{N} \tag{4}$$

4 Methodology

In this study, we validated the proposed metrics for creativity evaluation on three 100-image datasets, which were generated for the simulated task of object recognition (see Appendix A for dataset generation details). We performed statistical analysis to evaluate the difference in creativity measures between the three generated datasets. To enable statistical analysis, each of the three datasets was sub-sampled 100 times, taking a random sample of size 50 from each of the datasets and running it through each evaluator. Differences between dataset creativity scores (see Appendix B for programmatic creativity scoring details) were analyzed using standard statistical techniques: MANOVA, ANOVA, and t-tests. To test the validity of the novel evaluation method, we conducted a correlation analysis with all four dependent variables. Finally, we performed factor analysis with the hypothesis of a single underlying factor, as all four primary variables are meant to measure the same construct: creativity.

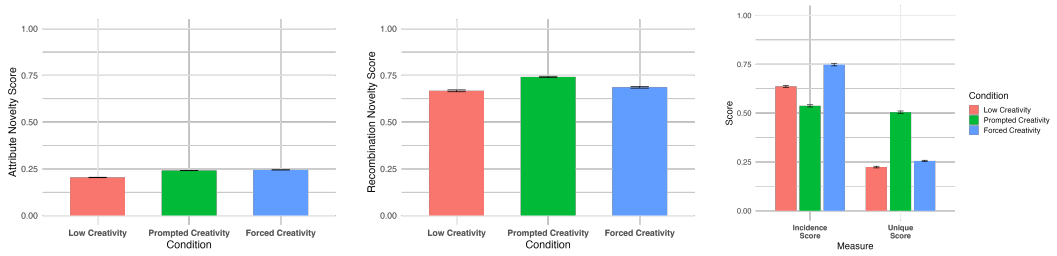
5 Results

Attribute novelty evaluation We found a significant multivariate effect of condition on the combined individual attribute novelty scores, Pillai's Trace = 1.656, $F(8, 590) = 354.96$, $p < .001$, $\eta_p^2 = .022$. There was also a significant effect of condition on average attribute novelty score: $F(2, 297) = 601.5$, $p < .001$, $\eta^2 = .802$ (see Figure 1a). Follow-up pair-wise t-tests indicated that the low creativity condition scored significantly lower than both the prompted and forced creativity conditions (both $p < .001$), and the prompted creativity condition was significantly lower than the forced creativity condition ($p = .011$).

Recombination novelty evaluation We found a significant effect of condition on recombination novelty score: $F(2, 297) = 76.35$, $p < .001$, $\eta^2 = .340$ (see Figure 1b). Follow-up pair-wise t-tests showed that prompted creativity had a significantly higher recombination novelty score than both the forced creativity and the low creativity conditions (both $p < .001$). Furthermore, the low creativity condition had a significantly lower recombination novelty score as compared to the forced creativity condition ($p = .0086$).

Feature addition We found a significant multivariate effect of condition on both feature addition outcomes, Pillai's Trace = 1.656, $F(8, 590) = 354.96$, $p < .001$, $\eta_p^2 = .022$. Significant univariate effects of condition were also observed where there was a significant effect of condition on unique score ($F(2, 297) = 1084$, $p < .001$, $\eta^2 = .880$) and a significant effect of condition on incidence score ($F(2, 297) = 313.5$, $p < .001$, $\eta_p^2 = .679$), see Figure 1c. For the incidence score, post-hoc t-tests showed that all conditions differed significantly, with forced creativity $>$ low creativity $>$ prompted creativity (all $p < .001$). Similarly, for the unique score, the three conditions differed significantly with prompted creativity $>$ forced creativity $>$ low creativity (all $p < .001$).

Inter-metric correlation Pearson correlations revealed significant positive associations between the following dependent variables: average attribute novelty score, recombination novelty score, and



(a) (b) (c)

Figure 1: Evaluation results measuring the effect of condition on: (a) the average attribute novelty score; (b) the recombination novelty score; (c) two feature addition measures.

unique score (all $r > .43$, all $p < .001$). Interestingly, the incidence score was negatively correlated with the unique score ($r = -.67$, $p < .001$) and with the novelty score of the recombination ($r = -.31$, $p < .001$) and had no significant association with the novelty score of the average attribute ($r = .07$, $p = .214$). Finally, a factor analysis was conducted on these four key variables, and the factor loading scores for average attribute novelty score, recombination novelty score, and the unique score were .43, .55, and 1.00, respectively, while the factor loading score for the incidence score was $-.67$.

6 Summary and operational implications

The proposed framework addresses critical challenges in generating training data for vision systems in ground vehicle recognition and overhead imagery analysis. The three-dimensional metric decomposition enables systematic and automated assessment of dataset creativity for object detection tasks in operational environments. For ground-based applications, attribute novelty ensures adequate environmental variability representation across terrain, lighting, and seasonal conditions that impact model robustness. Recombination novelty promotes diverse co-occurrence patterns among environmental attributes, preventing dataset saturation with common but incomplete combinations that can degrade real-world performance.

For overhead imagery, feature addition becomes critical when evaluating synthetic dataset completeness. The framework ensures appropriate distribution of infrastructure elements, vegetation patterns, and background objects that appear in operational imagery. The CLIP-based approach provides scalable diversity evaluation without human annotation, enabling rapid assessment before expensive training phases. This modality-agnostic methodology enables practitioners to identify specific deficiencies in synthetic data generation pipelines and systematically improve dataset quality for high-stakes deployment scenarios where model failure carries operational consequences.

References

[1] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3961–3969.

[2] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[3] S. Hossain and D.-J. Lee, "Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with gpu-based embedded devices," *Sensors*, vol. 19, no. 15, p. 3371, 2019.

[4] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, 2011, pp. 1521–1528.

[5] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman, "Dataset issues in object recognition," in *Toward Category-Level Object Recognition*. Berlin, Heidelberg: Springer, 2006, pp. 29–48.

[6] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?" *International Journal of Computer Vision*, vol. 119, no. 1, pp. 76–92, 2016.

[7] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017, pp. 746–753.

[8] D. H. Park, S. Azadi, X. Liu, T. Darrell, and A. Rohrbach, "Benchmark for compositional text-to-image synthesis," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1, 2021.

[9] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*, 2021, pp. 5637–5664.

[10] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Fidler, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in *2019 International Conference on Robotics and Automation*, 2019, pp. 7249–7255.

[11] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*, 2019, pp. 5389–5400.

[12] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler, "Meta-sim: Learning to generate synthetic datasets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4551–4560.

[13] E. Richardson, K. Goldberg, Y. Alaluf, and D. Cohen-Or, "Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints," *ACM Transactions on Graphics*, vol. 43, no. 3, pp. 1–14, 2024.

[14] A. Aghazadeh and A. Kovashka, "Cap: Evaluation of persuasive and creative image generation," *arXiv preprint arXiv:2412.10426*, 2024.

[15] G. Franceschelli and M. Musolesi, "Deepcreativity: measuring creativity with deep learning techniques," *Intelligenza Artificiale*, vol. 16, no. 2, pp. 151–163, 2022.

[16] A. Ramaswamy, H. Chockler, and M. Navaratnarajah, "Quantitative measures of task-oriented creativity in popular generative vision models," *arXiv preprint arXiv:2505.04497*, 2025.

[17] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[18] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" *arXiv preprint arXiv:1905.07830*, 2019.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[20] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, P. Poonam, T. Ropinski *et al.*, "A survey on quality metrics for text-to-image generation," *arXiv preprint arXiv:2403.11821*, 2024.

[21] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[22] N. Imasato, K. Miyazawa, T. Nagai, and T. Horii, "Creative agents: Simulating the systems model of creativity with generative agents," *arXiv preprint arXiv:2411.17065*, 2024.

[23] T. Shen, J. H. Liew, L. Mai, L. Qi, J. Feng, and J. Jia, "Empowering visual creativity: A vision-language assistant to image editing recommendations," *arXiv preprint arXiv:2406.00121*, 2024.

[24] Z. J. Hou, A. Kovashka, and X. L. Li, "Leveraging large models for evaluating novel content: A case study on advertisement creativity," *arXiv preprint arXiv:2503.00046*, 2025.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[26] M. L. Maher, "Evaluating creativity in humans, computers, and collectively intelligent systems," in *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*. Aarhus, Denmark: ACM, 2010, pp. 22–28.

[27] A. Shakhmatov, A. Razzhigaev, A. Nikolich, V. Arkhipkin, I. Pavlov, A. Kuznetsov, and D. Dimitrov, "kandinsky 2.2," 2023.

[28] P. Zhang, G. Zeng, T. Wang, and W. Lu, "Tinyllama: An open-source small language model," 2024.

[29] J. Li, D. Li, S. Savarese, and S. C. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[30] S. A. Research, "Salesforce/blip2-opt-2.7b," 2023, image captioning model hosted on Hugging Face. [Online]. Available: https://huggingface.co/Salesforce/blip2-opt-2.7b

[31] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spacy: Industrial-strength natural language processing in python," *Zenodo*, 2020.

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[33] S. A. Research, "Salesforce/blip2-flan-t5-xl," 2023, vision-language question answering model hosted on Hugging Face. [Online]. Available: https://huggingface.co/Salesforce/blip2-flan-t5-xl

[34] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

A Synthetic dataset generation

Three 100-image datasets were generated in this study to analyze our proposed measures of dataset creativity. Each of the three datasets was generated with different image generation prompts, leading to varying levels of creativity in each of the datasets. All of the images in the three datasets were generated using the Kandinsky 2.2 image generation model [27]. For the generation of each image, the Kandinsky model was initialized with the negative prompt of "low quality, bad quality," the prior guidance scale set to 2, and the generated image dimensions set to $900x900$ pixels.

A.1 Low creativity dataset

The low-creativity dataset was generated by passing pre-generated, low-creativity image generation prompts into the image generator. The image generator prompts contained two parts: the subject, which was always "An image of a silver-grey open-bed present-day pickup truck driving in" and the context, which varied by image. The image generation prompt contexts were generated using TinyLlama-1.1B [28], which followed text generation prompts meant to inhibit creativity. The creativity-inhibiting prompts for TinyLlama were as follows:

```
"role": "system",
"content": "You have the goal of creating concise scenery descriptions for
    image captions."

"role": "system",
"content": "You generate mountainous environment scenery descriptions
    varying season and weather conditions."

"role": "user",
"content": "Generate 10 concise, less than 20-word, scenery descriptions
    of mountainous environments. Include the word 'mountains'."
```

TinyLlama began to hallucinate when generating more than 10 contexts at a time, so contexts were generated in sets of 10. Only contexts from completed generation runs were included in the final dataset, and contexts containing nonsense text or failing to comply with the prompt were manually removed. 100 generated contexts were appended to the subject to create 100 image generation prompts. Example image generation prompts include "An image of a silver-grey open-bed present-day pickup truck driving in a snow-covered mountain ridge with a waterfall cascading down the rocks," "An image of a silver-grey open-bed present-day pickup truck driving in a winter wonderland with snow-capped mountains in the background," and "An image of a silver-grey open-bed present-day pickup truck driving in a deep, snow-covered valley surrounded by jagged peaks and the sky blue above." The 100 image generation prompts were passed into the Kandinsky model, and 100 images were generated. Images were manually screened and regenerated when the image generator failed to generate a semi-realistic image that adhered to the prompt.

A.2 Prompted creativity dataset

The prompted creativity dataset was generated in much the same way as the low creativity dataset. Again, the image generation prompts were split into a subject ("An image of a silver-grey open-bed present-day pickup truck driving in") and a context, which was again generated using TinyLlama. For the generation of this dataset, however, the text generation prompts passed into TinyLlama were designed to promote rather than inhibit creativity. The creativity-promoting prompts for TinyLlama were as follows:

```
"role": "system",
"content": "You have the goal of creating concise scenery descriptions for
    image captions."

"role": "system",
"content": "You generate scenery descriptions by varying biome, landforms,
    season, and weather conditions. You should also randomly include
    inconsequential details like animals, infrastructure, or plants."
```

```
"role": "user",
"content": "Generate 10 concise, less than 20-word, scenery descriptions
    for varying natural environments"
```

Following the same procedure as with the low creativity dataset, contexts were generated in sets of 10, and only contexts from complete generation runs were included. Also, contexts were manually inspected, and nonsense/noncompliant contexts were eliminated. Following the same procedure, contexts were appended to the subject to create 100 image generation prompts. Example prompts include "An image of a silver-grey open-bed present-day pickup truck driving in a vast open field, with rolling hills and patches of wildflowers, a herd of cattle grazing peacefully in the distance," "An image of a silver-grey open-bed present-day pickup truck driving in a dense jungle with towering trees and a lush undergrowth," and "An image of a silver-grey open-bed present-day pickup truck driving in a desert oasis with palm trees and crystal-clear water." The same image generation procedure was used, where the prompts were passed into the Kandinsky model, the generated images were manually screened to ensure realism, and select images were regenerated.

A.3 Forced creativity dataset

The forced creativity data set followed a unique procedure for image prompt generation. Instead of following a subject-context model where an LLM was used to generate the context, various context variables were predefined, and permutations of those variables were used to generate the different prompts. All of the image generation prompts followed the same general structure:

"An image of a silver-grey open-bed present-day pickup truck driving in [biome] environment with [landform] in the background in the [season] on a [weather] day. Including appropriate details like plants, animals, and buildings."

Lists of different biomes, landforms, times of day, and weather conditions were created, and permutations were used to generate image prompts. The four lists were as follows:

```
biomes = ["a beach", "a desert", "a forest", "a grassland", "a jungle",
    "an urban", "a rocky"]

landforms = ["a mountain", "a river", "an open sky", "hills", "a lake"]

season = ["fall", "winter", "spring", "summer"]

weather = ["windy", "rainy", "cloudy", "foggy", "partly cloudy", "snowy",
    "sunny"]
```

Of the 630 possible permutations, 100 permutations were randomly selected to be turned into image generation prompts. Example image generation prompts include "An image of a silver-grey open-bed present-day pickup truck driving in a rocky environment with a river in the background in the winter on a rainy day. Including appropriate details such as plants, animals, and buildings," "An image of a silver-gray open-bed present-day pickup truck driving in a desert environment with a river in the background in the spring on a sunny day. Including appropriate details such as plants, animals, and buildings," and "An image of a silver-grey open-bed present-day pickup truck driving in a jungle environment with a river in the background in the fall on a windy day. Including appropriate details such as plants, animals, and buildings." The same image generation procedure was used, where the prompts were passed into the Kandinsky model, the generated images were manually screened to ensure realism, and select images were regenerated.

A.4 Generation results

Three image datasets, each containing 100 images of trucks in natural environments, were successfully generated. Representative images from the low creativity, prompted creativity, and forced creativity datasets are included in Figures 2, 3, and 4, respectively. Most images in the low creativity dataset included mountains as prompted, but despite the context generation prompting, there was very little diversity in context, as most had snow on the ground with sun in the sky. The prompted creativity and forced creativity conditions had fairly similar images. Both sets of images had strong diversity in landform and ecosystem, but both fell short in weather/season diversity, both tending toward sunny

days regardless of the image prompt. The image generator used in this study is fairly outdated, and as a result, there are hallucinatory artifacts and nonsense geometry included in many generated images.

Figure 2: Representative examples from the low creativity dataset.

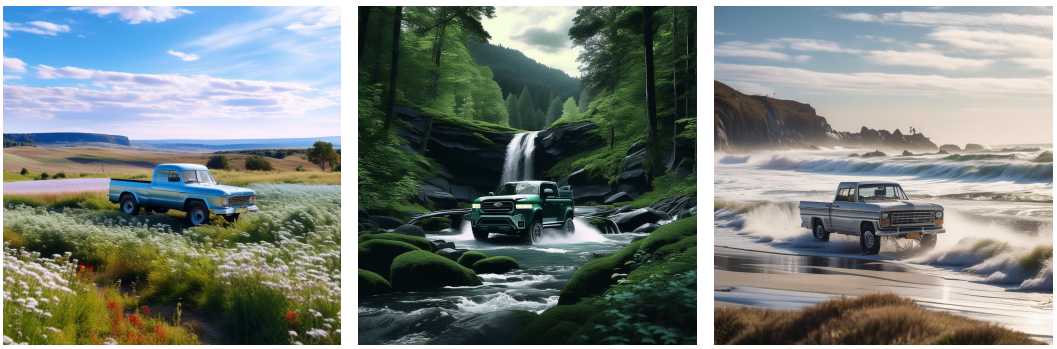


Figure 3: Representative examples from the prompted creativity dataset.

Figure 4: Representative examples from the forced creativity dataset.

B Creativity measures

B.1 Evaluator Initialization

Before running the evaluation, the creativity evaluator was initialized to reflect the downstream task of object recognition for pickup trucks in a variety of images of different natural environments. The universal attributes were set to "weather", "ecosystem", "landform", and "season" to reflect and capture the many different environments that the trucks will be depicted in. The unique attributes were set to "animal", "plant", and "infrastructure" to capture common image embellishments like small structures and plant/animal life. Both the universal and unique attributes are also explicitly prompted in the generation of the low creativity and the prompted creativity datasets. Finally, the evaluator included a new distractor attribute of "path", which captures artifacts like "road," "street," and "trail." These artifacts are referencing the surface that the depicted truck is driving on and should be excluded from analysis.

B.2 Artifact Extraction

To generate an analyzable list of artifacts for each image, every image in the dataset followed the same artifact extraction protocol. First, we employed BLIP-2 [29], using Salesforce's OPT-2.7b implementation [30] to generate a mid-length caption for each image with the maximum generation length set to 65 tokens and the minimum set to 55. Following caption generation, the noun phrases (artifacts) were extracted from each caption using spaCy [31] and categorized into one of the pre-defined attributes using BERT MNLI [32] with a corresponding probability score for that classification. For attributes like "season," which are unlikely to be included in a descriptive caption, Salesforce's Flan-T5-xl BLIP-2 implementation [33] was used for question answering, where the outcome of the query "Question: what is the season? Your options are winter, summer, spring, or fall. Answer with one word:" is used and labeled as a "season" artifact. Finally, CLIP [34] was used to calculate the semantic embedding, a 512-dimensional vector that represents the semantic meaning of the text, for each of the artifacts.

B.3 Attribute Novelty Metric

Attribute novelty was calculated for all three datasets at the same time, following Eq. 1 described in Section 3. First, the artifact sets were cleaned so all artifacts were eliminated except for those with the highest probability scores for each attribute in each image. Then, a list of artifacts was created for each universal attribute, where the list contained all of the artifacts categorized as the given universal attribute across the entire dataset. To ensure equal-sized comparisons between datasets, the datasets were sampled 100 times, where the sub-sample size was equal to half of the size of the smallest artifact list between the three datasets. Following the calculation of the sub-sample size for each of the universal attributes, the attribute novelty score was calculated for each of the unique attributes individually. The attribute novelty score was calculated by taking the average distance between the semantic embeddings of the artifacts in each sub-sampled dataset and averaging over all sub-sampled datasets. Finally, the total attribute novelty score was calculated by averaging over each of the unique attribute novelty scores.

B.4 Recombination Novelty Metric

Recombination novelty was calculated for each dataset individually following Eq. 2 in Section 3. Again, the datasets were cleaned so only the artifacts with the highest probability score in each classification category are retained. Then the embeddings which each attribute category were clustered using agglomerative clustering, where the optimal number of clusters was determined according to the silhouette score using cosine similarity as the similarity metric.

After all of the artifacts across every attribute classification were given cluster groups, the novelty of recombination was calculated. A list was generated of all the possible pairs of universal attributes, and for each pair, all of the pair instances were identified and counted across the dataset. Then, using the clustering group numbers, the total number of unique pairs was calculated. This was done for each of the pairs, and the average ratio of unique pairs was returned as the recombination novelty score.

B.5 Feature Addition Metric

Recombination novelty was calculated for each dataset individually following Eq. 3 and 4 in Section 3. A similar agglomerative clustering method was employed; however, a threshold rule was used instead of an optimization rule in determining the total number of clusters. The embeddings for each of the unique attributes were clustered with a cosine similarity threshold of .85. The uniqueness of added features was calculated as the ratio of unique clusters to total added features, and the frequency of added features was calculated as the number of images with artifacts that were classified as unique attributes.