Defining and Quantifying Creative Behavior in Popular Image Generators

Aditi Ramaswamy
aditi.ramaswamy@kcl.ac.uk
and **Hana Chockler**
hana.chockler@kcl.ac.uk
and **Melane Navaratnarajah**
melane.navaratnarajah@kcl.ac.uk

Abstract

Whether generative AI models can exhibit creative behavior has been a subject of scientific debate in the last years, without a conclusive answer. In this paper, we define creativity from a task-oriented perspective and introduce quantitative measures that help the user to choose a suitable AI model for a given task. We have evaluated our measures on chains of images iteratively generated using number of popular `img2img` models. We observed that the images produced in these chains matches up with our metrics, with higher-scoring images exhibiting visual traits more strongly aligned with our definition of task-oriented creativity.

Introduction

Generative AI has become the center of a series of heated debates, one of the biggest ones centered around the concept of human creativity. Some researchers argue that creativity is derived from an innately human "sociocultural context", thus excluding AI models (Oppenlaender 2022; Kaufman, Sternberg, and Glăveanu 2019; Wingström, Hautala, and Lundman 2022). However, other definitions focus on purely behavioral requirements, such as "domain-relevant skills", "creativity-relevant processes", and "extrinsic motivation", all of which can be exhibited by generative AI models (Amabile 1983). Related to this, Margaret Boden proposes an AI-relevant cognitive dimension of creativity separate from emotion, focusing on behaviors such as novelty and association of ideas (Boden 1998). Boden also discusses "exploratory" creativity, the "generation of novel ideas" within a space, such as an AI model responding to a given task by creating new content within the parameters of the task. Even more broadly, Pease and Colton argue that computational creativity should be separated from human-centric notions, and creative behavior in AI models should be measured in entirely separate ways from human creativity (Pease and Colton 2011). These process-oriented definitions suggest that some generative models can exhibit a form of *task-oriented* creative behavior in response to prompts, thus forming the motivation for this paper.

Quantifying this task-oriented creative behavior could provide valuable insights for both the users who interact with these models and the model developers. If a user is presented with clear information about the ways in which a given generative model may exhibit creative behavior in response to their queries, they can make an informed choice about whether they wish to use it. The same information could also pinpoint specific areas of change that those who develop generative models can use to tweak their training data and model architecture to better suit their intended purpose.

The focus of this paper is answering the following research question:

"Can we define mathematical measures quantifying useful aspects of task-oriented creative behavior of popular `img2img` generation models?"

We propose to characterize the *task-oriented creativity* of model outputs according to three criteria, which we derived from prior literature on computational creativity as discussed in our *Background and Definitions* section. These criteria are: (1) satisfaction of prompt requirements, (2) cohesion between output artifacts, and (3) novelty of output artifacts. In our *Methodology* section, we introduce mathematical metrics and an iterative process involving construction of image *chains* to measure these, and we subsequently analyze the results through both statistical and visual means. The chain construction is inspired by the *Telephone* game and includes a repeated generation of new outputs based on the previous ones. This process is repeated for a predefined number of steps, with the resulting chain analyzed for the concepts introduced above. To the best of our knowledge, while previous papers have described aspects of creative behavior, this is the first attempt at mathematically measuring useful aspects of task-oriented creativity in popular `img2img` generation models.

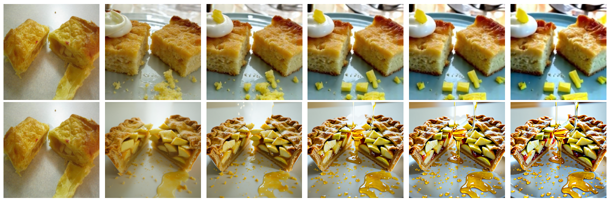

Figure 1: KANDINSKY 2.2 and STABLE-DIFFUSION 3 chains seeded with the same image, donuts_014.

Due to the brevity of this paper, we only present the summary results and a very small number of illustrative images from our experiments. The code, full experimental results, and chains of images can be downloaded from `https://figshare.com/s/d88d4966b606163d02fc`.

Background and Definitions

We use the term "artifact", derived from (Hauhio 2024)'s usage as well as the common linguistic meaning, to refer to a core feature of a text or image, such as "apple pie" in the textual prompt "a slice of apple pie", or a segment of an image labeled "pie" by an object detection tool. Each prompt and output can therefore be understood as a set of artifacts.

Hauhio's paper proposes multiple artifact spaces, the relationship between which can be used to classify models' behavior as creative. C, referred to as the "conceptual set" by (Hauhio 2024), is the set of artifacts specified in the user's input prompt to a generative model, and V is the set of artifacts a user would find valuable given that this user has a specific goal in using the generative model. Expanding the set of valuable artifacts from strictly C to encompass distinct ones from V as well can therefore be understood as one facet of creative behavior.

We propose to characterize outputs of models according to three criteria, whose formal quantitative definitions we present below: input prompt *requirement satisfaction*, *cohesion*, and *novelty*.

Satisfaction of Prompt Requirements, the idea that an output contains (all) the elements specified in the user's input, builds off (Peeperkorn et al. 2024)'s statement that an output must be typical of its class, as well as (Hauhio 2024)'s "concept set" C. We build *cohesion* from (Peeperkorn et al. 2024)'s assertion that an output must also be *useful* to be considered creative, by working off the idea that an output with elements that are generally more closely related to each other has more use cases than an output with disjoint elements. *Cohesion* is also derived from (Boden 1998)'s "association of ideas": how well does a generative model build strong associations between artifacts within its output in response to a given prompt? For *novelty*, we build off (Hauhio 2024)'s idea of valuable artifacts that were not specified by the input prompt, but are added spontaneously by the generative model, as well as (Peeperkorn et al. 2024)'s and (Boden 1998)'s ideas that an output must contain novel elements to be considered creative. Specifically, (Boden 1998) mentions "transformational" creativity as the idea that new elements can be introduced into a space through transformation, a behavior directly linked to the ability of `img2img` models to generate novel output elements using the framework of an image input.

These measures form a system of checks and balances. Novelty is widely accepted as a necessity for creative output, but including *requirement satisfaction* as a core aspect of task-oriented creativity ensures the penalization of image generators that hallucinate and do not follow task requirements, and *cohesion* ensures that generators are rewarded for producing novel outputs that are semantically comprehensible. This balance aligns with prior ideas surrounding computational creativity, such as Pearce and Wiggins' finding that human reviewers reward AI-generated musical compositions that fall within a range of novelty that does not strike the reviewers as "too strange" (Pearce and Wiggins 2001).

In the definitions below we use the following notations: M is the generative model, A_O is the artifact set of a given output image O_M (generated by M), and A_I is the artifact set of the input prompt I, which can be an image or a text. The *cosine similarity* measure, which is used throughout these definitions, is a way of measuring the similarity between the vector projections of two artifacts in semantic embedding space—in other words, a way to mathematically compare the meanings behind textual representations of artifacts. It will be henceforth denoted as the function θ or as the dot product $\cdot$, depending on context. For a set A of artifacts, the embedding of each artifact $a \in A$ in semantic space is calculated using the embedding function emb, denoted as $emb(a) = \hat{a}$ (Matsuki, Lago, and Inoue 2019).

Definition 1 (*Satisfaction of Prompt Requirements*) *An image is said to be* satisfying requirements *if $A_I \subseteq A_O$. In other words, all artifacts in I are present in O_M.*

To measure the cohesion of O_M, we wish to capture the sentiment of "how much do any artifacts introduced by the generator make sense within the context of the prompt?" To accomplish this, we must measure the closeness of the relationships between seed artifacts and introduced artifacts in the generated image without being too dependent on our third measure, *novelty*, while also not penalizing the generator for low similarity between input prompt elements. We first calculate the mean of the semantic similarity scores between each artifact of A_I and the closest artifact from A_O, and then calculate the maximal similarity between artifacts from A_O. We do this in order to algorithmically approximate the relationships between elements in the generated image without being too dependent on the cardinality of the set of introduced elements, while also not penalizing the generator for low similarity between input prompt elements. As a hypothetical illustrative example, if the input prompt is "a potato and a lion", then $A_I = potato, lion$. If, in the subsequent generated image, $A_O = potato, lion, onion, R2D2, zebra, strawberry$, and each element is represented by its first initial, cohesion can be calculated by first taking $\hat{P} \cdot \hat{O},\ \hat{L} \cdot \hat{Z},\ and\ \hat{O} \cdot \hat{S}$, then taking the average of those three values. This is because the input artifact "potato" has the highest semantic similarity to "onion" within $A_O \setminus A_I$, the input artifact "lion" has the highest semantic similarity to "zebra" within $A_O \setminus A_I$, and the two artifacts within $A_O \setminus A_I$ with the highest semantic similarity are "onion" and "strawberry". Note that "R2D2" is excluded from these measurements because only the maximally similar pairs are considered, and "R2D2" is not maximally similar to either of the seed artifacts, nor is it maximally similar to any of the other introduced artifacts.

Definition 2 (*Cohesion*) Cohesiveness $C(O_M)$ *of an output image O_M is defined wrt the input I. Given the set A_I of the artifacts in I and the set A_O of the artifacts in O_M, let P be the set of all pairs (a, b), where $a \in A_O \setminus A_I$ and $b \in A_I$. Furthermore, let $M_N = \max_{(c,d) \in A_O \setminus A_I}((\hat{c}) \cdot (\hat{d}))$. Then, cohesiveness of O_M is defined as*

$$C(O_M) = \frac{\sum_{a:(a,b) \in P} \max_{b:(a,b) \in P}((\hat{a}) \cdot (\hat{b})) + M_N}{|P| + 1}.$$

We note that the scalar product $(\hat{a}) \cdot (\hat{b})$ of the representations of a and b in the semantic embedded space increases with increasing similarity between a and b. As $0 \leq (\hat{a}) \cdot (\hat{b}) \leq 1$, also $0 \leq C(O_M) \leq 1$.

Definition 3 (*Novelty*) Novelty $D(O_M)$ *of an output image* O_M *is defined as the proportion of new artifacts in the image, if* O_M *contains all artifacts of* I. *In other words,* $D(O_M)$ *is* $\frac{|A_O \setminus A_I|}{|A_O|}$ *if* $A_I \subseteq A_O$ *and is* 0 *otherwise.*

While *novelty* and *cohesion* are not independent concepts, an output image O_M can have high *novelty* and low *cohesion* if it contains many new artifacts with low similarity to the original ones, and it can also have high *cohesion* and low *novelty* if there are very few new artifacts and they are very similar to the original ones.

Methodology

In this section, we outline our methodology for quantitative evaluation of task-oriented creativity in image generation models.

Chain Construction

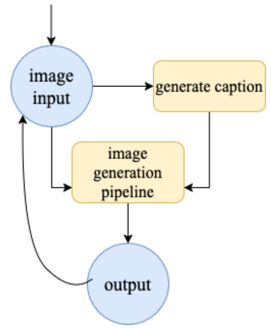

Figure 2: One step in the iterative chain process.

As (Hauhio 2024) points out, image generation models are often used iteratively, with users modifying their input prompts based on the artifacts present in the generated output. In this paper, we simulate this iterative approach, which allows us to measure the afore-discussed aspects of creative behavior in a way that is more pertinent to real-world use-cases. Similar automated approaches have been used for related tasks such as image generation refinement (Wang et al. 2025), but to our knowledge this is the first time such an iterative process has specifically been used to quantify any form of creative behavior exhibited by AI models.

Our experimental setup evaluates how strongly both text and image inputs influence aspects of task-oriented creativity as exhibited by generative models. We base our approach on the well-known game *telephone* (or *Chinese whispers*). As in this game, each chain in our experiment begins with a "ground truth", or seed image. For each step k in the chain, we use the image generated in step $k-1$ as input to the generative model. To produce the text input, we auto-generate a caption for the generated image from step $k-1$ and feed that into the image generation model alongside image $k-1$. This process repeats for x iterations, where x is a parameter chosen by the user. Figure 2 illustrates this process.

Measures for Creativity

Using the conceptual definitions of *requirement satisfaction*, *cohesion*, and *novelty* provided earlier, we can define mathematical measures for each of these aspects in the context of the chains, so that we can quantify the creative behaviors exhibited by the generator used within a given chain.

The *satisfaction factor* measure provides the backbone for the other two, as we want to focus on images that exhibit a higher-than-random level of *requirement satisfaction* for the input prompt. We expand and quantify Definition 1 to produce this measure for a chain of images rather than a single comparison between two images, by measuring how well the prompt requirements are propagated throughout a chain. Therefore, for a given chain, we quantify *satisfaction factor* (RS) as the *normalized longest unbroken sequence* of the chain wherein, at each step, the seed artifact set, or a set of artifacts that approximate the seed artifact set within a cosine similarity threshold t, is a subset of the artifact set of the image generated for that step.

We then multiply this proportion of the chain by the average θ score for the set of pairs consisting of one artifact from the seed set and its closest match in the generated image's artifact set, to account for the fact that the generated image may not be perfectly faithful.

The later in the chain the seed artifacts or their approximations still appear, the higher the *satisfaction factor* value is going to be, although it also depends on how closely the artifacts in the generated image approximate the chain image.

A chain's *Cohesion Factor* (B_R) and *Novelty Factor* (D_R) can be calculated using Definitions 2 and 3 from the last image in the unbroken chain as described in the *Requirements Satisfaction Factor* metric.

An example illustrating the calculation of these three metrics follows: Given a hypothetical chain generated from a seed artifact set ["apple pie" [AP], "horse" [H]], we observe that the eighth generated image in the chain is the last one that falls within the chosen cosine similarity threshold. Furthermore, we observe that has the artifact set ["apple cake" [AC], "horse" [H], "pear" [P], "jockey" [J]]. The proportion of the chain that satisfies the input prompt would be 0.8, and to calculate the *Requirements Satisfaction Factor*, this value would be multiplied by the average similarity scores between the closest corresponding elements from both sets, in this case AP, AC and H, H:

$$\frac{\theta(AP, AC) + \theta(H, H)}{2}.$$

The *Cohesion Factor* would be calculated as

$$\frac{\theta(AP, P) + \theta(H, J) + \theta(P, J)}{3}.$$

The *Novelty Factor* would be 0.5, given that half the elements of the generated image artifact set are not present in the input artifact set.

In order to measure overall task-oriented creativity, we combine these measures to an *overall task-oriented creativity* *ranking* score, CR, where $0 \leq CR \leq 1$. For a given chain:

$$CR = RS \times \frac{B_R + D_R}{2}. \quad (1)$$

The chain under evaluation is a parameter to all scores we introduce in this section; it is omitted for brevity.

Intuitively, CR can be higher than 0 only if the generated images satisfy the prompt requirements; this prevents us from awarding a high CR to hallucinating models. On the other hand, a mechanistic copying of the seed image is not creative either. Indeed, while $RS = 1$ for a mechanistic copy, both B_R and D_R are 0, hence resulting in the creativity score 0 as well. For models that satisfy the prompt requirements and add new elements to the generated images, both the *novelty* between the new elements and their cohesion with the seed image elements, as well as between themselves, affect CR.

Given a set X of generative models, we can use the means of RS, B_R, D_R, and CR across the total number of chains for all models in X to rank them in terms of specific aspects of task-oriented creative behavior as per our earlier discussion. Of course, this ranking does not necessarily correlate with general creativity. Extensions or modifications to these definitions can be explored to better fit specific domains or to align with different interpretations of task-oriented creativity.

Evaluation

Experimental Setup

For our experiments we selected sets of simple seed images focusing on a single, easy-to-detect subject, as these would be the easiest to measure *requirement satisfaction*, *cohesion*, and *novelty*, since A_I would be 1. Since many object detectors are trained on datasets that prominently feature food images, we pulled our 999 seed images from the publicly-available `Food-101` dataset (Bossard, Guillaumin, and Gool 2014). We selected three food categories at random: "apple pie", "donuts", and "pizza", and randomly selected 333 images from each. Input prompts I were constructed from each of these seed images.

We evaluated our measures on three popular open-source image generation models from Hugging Face: STABLE-DIFFUSION 3 3 (8.1 billion parameters) (Esser et al. 2024), KANDINSKY 2.2 (4.6 billion parameters) (Razzhigaev et al. 2023), and FLUX (12 billion parameters) (BlackForestLabs 2023). For the text input, we generate a text caption for the seed image using the KOSMOS model from Hugging Face (Peng et al. 2023), and the image input consists of the image generated in the previous step of the chain (or, for the first generative step, an image from the seed dataset). In our experiments, we set the maximal length of the chain to 10 to avoid computational explosion, while still producing meaningful results.

To compute the measures of *requirement satisfaction*, *cohesion*, *novelty*, and creativity ranking defined earlier, we used two object detectors, GROUNDINGDINO (Liu et al. 2023) and DETR (Carion et al. 2020), to extract artifacts in textual form from images in the chain, where the initial image is the seed image, and for A_I we simply use a text representation of the subject of the seed image: apple pie, donut, or pizza. For word embedding and semantic similarity calculations between the artifacts, we used the SPACY library (Montani et al. 2023). We selected 0.65 as the value of t for *satisfaction factor*, after running preliminary tests to account for incorrect object detection results.

We used *paired T-tests*, which are statistical tests tailored to compare the application of different processes to the same input data points, in order to determine any statistically significant differences between results for different image generation models, on the same dataset of 999 images. We used SCIPY's `ttest_rel` function to perform these tests.

To verify reproducibility, we repeated a subset of the experiment with 100 randomly picked seed images, confirming that the statistical differences between the reruns match the ones we extracted from the main experiment, with any discrepancies in averages being explained by the vastly reduced size of the seed dataset in the rerun.

Analysis

Our first observation is that the average CR across all the models was relatively low, ranging from 0.08 to 0.18. This means that none of the image generation models we tested performed extremely well in terms of task-oriented creativity. Manual examination showed that these low scores were due to multiple factors: early deviation from the prompt requirements, satisfaction of prompt requirements but low *novelty*, or the generation of unclassifiable objects such as glitchy patterns or irregular geometric shapes. However, the paired T-tests we ran support the claim that task-oriented creativity is measurably different between some generative models. We noted that FLUX displayed a statistically significantly higher average *Requirements Satisfaction Factor* (0.74) than the other two models (0.7 for KANDINSKY 2.2 and 0.45 for STABLE-DIFFUSION 3), but FLUX and KANDINSKY 2.2 performed about the same on *Cohesion Factor* (0.15), *Novelty Factor* (0.25 vs 0.24 respectively), and overall task-oriented creativity CR (0.18 vs 0.17 respectively). STABLE-DIFFUSION 3, however, displayed statistically significantly lower averages across the board, indicating a much lower performance on task-oriented creativity measures than FLUX and KANDINSKY 2.2.

Two sample chains from our process are illustrated in Figure 1. The top chain of images was produced using the model FLUX, while the bottom chain was produced using the model STABLE-DIFFUSION 3. The leftmost image in both chains is the non-generated seed image, with $A_I = apple\ pie$, while the rest were generated by the respective models. In the FLUX chain, *requirement satisfaction* is maintained through the generated images, which contain $A_O = cake$, an artifact that falls within the cosine similarity threshold t as described in our *Methodology* subsection on *Measures for Creativity*. The model also adds novel artifacts: "plate", "cream", "crumbs", and "fork" (with the latter being out of focus in the background of the image). All of these introduced artifacts exhibit high cohesion, since they make sense in the context of the seed image and are all generally highly semantically related: "crumbs", for example, have a close semantic relationship to "apple pie", which also has a close semantic relationship to items like "cream" and "plate". The STABLE-DIFFUSION 3 chain, seeded with the same image, maintains *requirement satisfaction* through propagating the seed artifact “apple pie” throughout the chain, and adds two novel elements: “syrup” dripping down onto the pie, and “crumbs” surrounding the pie.

Notably, the object detection tools used within this experiment failed to detect any of the extraneous artifacts, an issue that is unavoidable when using automated tools, but manually calculating the scores for the first chain, dubbed C_1, yields the following values according to our mathematical definitions as per our *Methodology* subsection on *Measures for Creativity*: $RS(C_1) = 0.68$, $B_R(C_1) = 0.52$, and $D_R(C_1) = 0.8$. For the second chain, dubbed C_2, the calculated values are $RS(C_2) = 1.0$, $B_R(C_2) = 0.52$, and $D_R(C_2) = 0.67$. Based on these values, $CR(C_1) = 0.45$ and $CR(C_2) = 0.6$, with the second chain scoring higher due to preserving the seed artifact perfectly. Both chains display medium levels of task-oriented creativity, with some distinctively novel elements. Note that SPACY is still used to calculate *cohesion*, with the maximally similar pairs for C_1 being “apple pie”, “cream” and “crumbs”, “cream”, and the maximally similar pairs for C_2 being “apple pie”, “syrup” and “crumbs”, “syrup”.

Limitations

Apart from computational resources, a significant limitation of our approach is its reliance on existing popular object detection and word similarity calculation tools. These models are prone to making decisions that do not seem intuitive to humans, such as placing two seemingly unrelated words close together in semantic embedding space, labeling one object with a similar but incorrect designation (for example, labeling a slice of apple pie as a slice of cake), or failing to detect an object altogether, as described in our prior discussion of Figure 1. There is currently no way around this limitation, but as the outputs of such tools more closely approach human-level understanding, the quality of our measures will improve accordingly. Additionally, in order to improve the visual accuracy and explainability of our proposed measures, future research could incorporate human observation as a factor in validating the scores output.

We also acknowledge that the simplicity of our seed image dataset, although easy to experiment on and ideal for observing changes, could impact creativity scores, especially since we set the seed artifact sets to only contain the main subject of the image, ignoring extraneous elements in the seed images (such as plates, tables, and phones) for clarity. Seed images with a large artifact set may potentially increase D_R, for example, or make it harder for generated images to satisfy prompt requirements. Exploring task-oriented creativity across diverse domains and more complex datasets subjects is a crucial next step, particularly since the metrics defined within this paper are meant to accommodate that level of complexity.

Conclusions and Future Research

In this paper, we suggest a set of measures to quantitatively evaluate task-oriented creativity in generative models, and run a series of experiments to study their manifestation in a set of popular `img2img` models. The results of statistically comparing these metrics show that some models perform higher on task-oriented creativity measures than others, and this is supported by visual examination of the results.

We can, therefore, answer the research question posed in the introduction affirmatively: our metrics provide a fine-grained and intuitive score of creative behavior in generative models when responding to a prompt, which in turn permits users a greater ability to select the appropriate model for a given task. Further directions for this research include exploring more models, explaining why some generative models score differently, and determining whether these metrics can be generalized to tailor future model architecture and training techniques to specific purposes.

References

Amabile, T. M. 1983. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology* 45(2):357–376.

BlackForestLabs. 2023. Flux. `https://github.com/black-forest-labs/flux`.

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103(1):347–356. Artificial Intelligence 40 years later.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 – mining discriminative components with random forests. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 446–461. Cham: Springer International Publishing.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Muller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv* abs/2403.03206.

Hauhio, I. 2024. Enhancing human creativity with aptly uncontrollable generative ai. In *Proceedings of the 15th International Conference on Computational Creativity*. Spain: Association for Computational Creativity (ACC). International Conference on Computational Creativity, ICCC ; Conference date: 17-06-2024 Through 21-06-2024.

Kaufman, J. C.; Sternberg, R. J.; and Glăveanu, V. P. 2019. *A Review of Creativity Theories*. Cambridge University Press. 27–43.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Matsuki, M.; Lago, P.; and Inoue, S. 2019. Characterizing word embeddings for zero-shot sensor-based human activity recognition.

Montani, I.; Honnibal, M.; Honnibal, M.; Boyd, A.; Van Landeghem, S.; and Peters, H. 2023. explosion/spacy: v3.7.2: Fixes for APIs and requirements.

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

Pearce, M., and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, 22–32. AISB (Society for the Study of Artificial Intelligence and the Simulation of Behaviour).

Pease, A., and Colton, S. 2011. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *AISB 2011: Computing and Philosophy*.

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *International Conference on Computational Creativity*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv* abs/2306.14824.

Razzhigaev, A.; Shakhmatov, A.; Maltseva, A.; Arkhipkin, V.; Pavlov, I.; Ryabov, I.; Kuts, A.; Panchenko, A.; Kuznetsov, A.; and Dimitrov, D. 2023. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In Feng, Y., and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 286–295. Association for Computational Linguistics.

Wang, J.; He, Y.; Zhong, Y.; Song, X.; Su, J.; Feng, Y.; He, H.; Zhu, W.; Yuan, X.; Lu, K.; Huo, M.; Zhang, M.; Li, K.; Chen, J.; Shi, T.; and Wang, X. 2025. Twin co-adaptive dialogue for progressive image generation. *arXiv*.

Wingström, R.; Hautala, J.; and Lundman, R. 2022. Redefining creativity in the era of ai? perspectives of computer scientists and new media artists. *Creativity Research Journal* 36(2):177–193.