

Design Guidelines for Prompt Engineering Text-to-Image Generative Models

VIVIAN LIU, Columbia University, USA

LYDIA B. CHILTON, Columbia University, USA

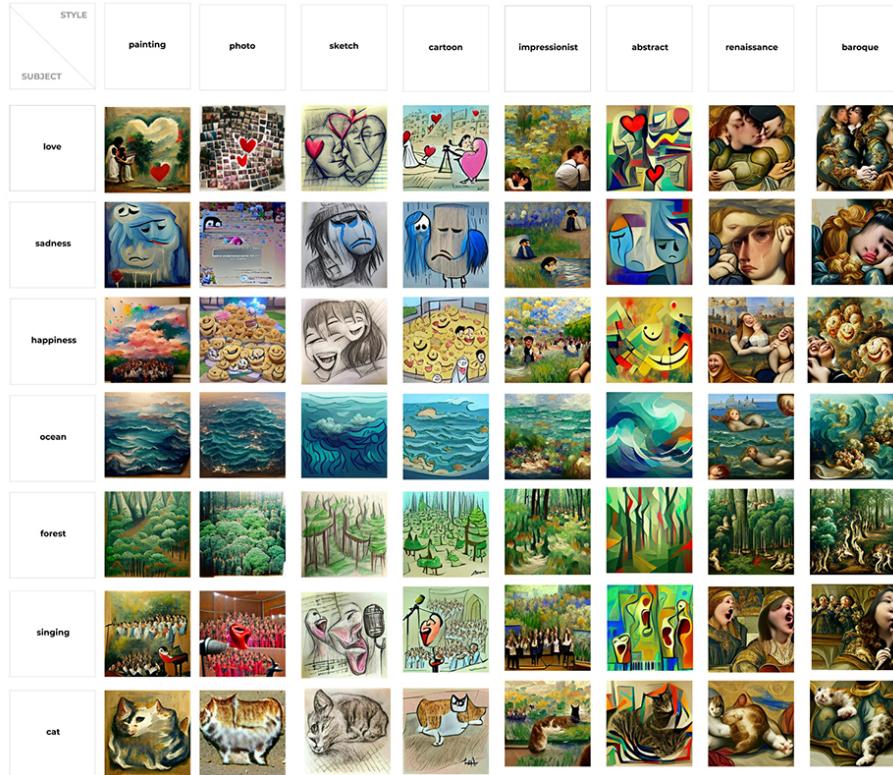


Fig. 1. A series of five prompt engineering experiments structured using the template "SUBJECT in the style of STYLE" was conducted across 49 subjects and 51 styles of varying levels of abstractness.

Text-to-image generative models are a new and powerful way to generate visual artwork. The free-form nature of text as interaction is double-edged; while users have access to an infinite range of generations, they also must engage in brute-force trial and error with the text prompt when the result quality is poor. We conduct a study exploring what prompt components and model parameters can help produce coherent outputs. In particular, we study prompts structured to include subject and style and investigate success and failure modes within these dimensions. Our evaluation of 5493 generations over the course of five experiments spans 49 abstract and concrete

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

subjects as well as 51 abstract and figurative styles. From this evaluation, we present design guidelines that can help people find better outcomes from text-to-image generative models.

ACM Reference Format:

Vivian Liu and Lydia B. Chilton. 2021. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of ACM, New York, NY, USA, 27 pages.*

1 INTRODUCTION

Recently, advances in computer vision have introduced methods that are remarkable at generating images based upon text prompts. [27, 28]. For example, OpenAI introduced DALL-E, one such text-to-image model in 2020, and demonstrated that from running a text prompt such as "*a radish dancing in a tutu*", the model could generate many images matching the prompt. Based on this progress, artists, programmers, and researchers have come together on communities within Reddit and Twitter like r/bigsleep and #vqganclip to develop different models to open source text-to-image generation. [1, 3] Tutorials [2, 10] and interactive notebooks [30] maintained by community members such as @RiverHaveWings, @advadnoun, and @somewheresy on Twitter have made these tools broadly accessible. [4, 9, 20]

Text is free-form and open-ended and the possibilities for image generation are endless. However, this also means that the design process for generating an image can become brute-force trial and error. People must search for a new text prompt each time they want to iterate upon their generation, a process that can feel random and unprincipled. In the field of natural language processing, this problem known as *prompt engineering* [29]. Prompt engineering is the formal search for prompts that retrieve desired outcomes from language models, where what is desirable is dependent upon the end user. There are a number of open questions within prompt engineering to explore for text-to-image models. Some questions relate to hyperparameters: how do variables influencing the length of optimization and randomness affect model outcomes? Other questions revolve around the prompt: are there certain classes of words or sentence phrasings that yield better outcomes?

In this paper, we undertake a systematic investigation of many of the variables involved in the prompt engineering process. Over the course of five experiments, we test different parameters and methods that can help users navigate the space of text-to-image design solutions and steer themselves towards desirable generations. In terms of parameters, we look at different permutations of prompt keywords, the influence of random seeds, and the length of iterations. In terms of methods, we investigate whether prompt can be structured using *subject* and *style* as dimensions. We further analyze how generation quality is impacted by how abstract a subject or style can be. From these experiments, we present a series of design guidelines to help users prompt text-to-image models for better outcomes.

2 RELATED WORK

2.1 Generative Methods as Creativity Support Tools

Artist and programmer communities have shown a consistent interest in the potential of generative AI as an art medium. Communities conversing about artistic AI have developed for a long time around techniques, algorithms, and networks such as DeepDream [26], neural style transfer [11], and generative adversarial networks. [6, 16, 21] Likewise, HCI researchers have sought to understand how generative AI can become a creativity support tool for artists. In recent years, systems embedding generative models have been successfully applied to domains such as image generation, poetry, and music.[14, 15, 18, 19]

Often, researchers want to use the large space of design solutions of these generative methods to assist users in ideation and divergent thinking. [24] Thus, an important angle of research around generative models is how we should let users explore their space of design solutions efficiently and effectively. This question has been investigated under a number of computational approaches. For example, Matejka et. al. introduced Dreams Lens, a system that utilized design galleries with interactive data visualization to summarize the diversity of design solutions from a generative program for 3D modeling. [25] Yang et al. proposed a domain-agnostic method known as latent space cartography, which leveraged dimensionality reduction to explore the latent space of generative models. [23]

One drawback of many of these methods exploration is that they lack meaningful controls. This has given rise to a recent trajectory within the intersection of HCI and AI focused on semantically meaningful exploration. One of the earliest works in this direction was a seminal creativity support system called AttriBit, which allowed users to assemble 3D models given data-driven suggestions. [8] They prompted the model with semantic goals they had for their creations. More recent work leveraging deep learning has tried to produce semantically meaningful editing operations. For example, Louie et. al. introduced in their novice music creation system CoCoCo a feature that allowed users to use sliders affecting the mood of the music piece by editing it to be "happier" or "sadder". [24]

2.2 Text-to-Image Generation

This interest in involving natural language as a form of interaction with generative models has recently found success in computer vision. Recent work from computer vision has focused on learning text and image together in generative models, by coupling the two modalities through optimization. In 2021, Radford et. al. from OpenAI introduced CLIP (Contrastive Language-Image Pre-training), a method for learning multimodal image representations. [27] CLIP was trained on an Internet scale size dataset of 400 million image and text pairs to optimize a contrastive objective that pushed image and text embeddings closer towards one another. CLIP demonstrated that the model was able to learn "visual concepts...enabling zero-shot transfer of the model" on various tasks such as OCR, geo-localization, and others. CLIP was used in DALL-E [28], one of the state of the art models for text-to-image generation. DALL-E learned a transformer that autoregressively predicted text and image tokens together in one sequence. The authors of DALL-E demonstrated how the model could handle image operations, perform style transfer, and produce novel combinations of elements.

2.3 Prompt Engineering

Researchers have begun to explore the space of prompt engineering for text-to-image generative models. Ge and Parikh et. al. [12] utilized text-to-image models to generate visual blends of concepts. BERT, a large language model, was then used to help users generate prompts, and the generations were evaluated using crowd-source workers on Mechanical Turk. Similar large-scale crowd-sourced approaches have been done on the past to evaluate machine-generated images for the quality and coherency of generations. [5]

The term prompt engineering was originally reified in a long-form post online by a Gwern, a writer and technologist who evaluated GPT3's capabilities on creative fiction. [17] Gwern positioned prompt engineering as an emerging form of interaction with generative models that have learned complex abstractions from consuming Internet-scale quantities of data. Because these large models have metalearning capabilities (they have learned how to learn), they can adapt their abstractions on the fly in the form of soft attention weights to fit new tasks. Gwern suggested that a new course of interaction would be to figure out how to prompt the model to elicit the specific knowledge and abstractions needed to perform well on new tasks.

Prompt engineering is a valuable form of interaction to study in detail because it can help users develop mental models of the generative models. In a paper investigating human-AI interactions in a game play setting, Gero et. al. demonstrated that people can construct mental models of AI through repeated interactions with a model that help them understand the AI's knowledge distributions, local behavior, and global behavior.[13] Using prompts to generate image evidence of AI knowledge is also a way of reducing uncertainty with AI, one of the fundamental challenges of human-AI interaction.[31]

On the Internet, hobbyists and researchers have already been engaging in prompt engineering as well. They have been discussing "tricks" and keywords to understand the knowledge distribution of the text-to-image models and to discover what keywords can help tune models towards their aesthetic goals. One prominent artist and research programmer noted that using 'unreal engine' as a prompt helped them add a hyperrealistic, 3D render quality to their image generation. This tweet and many others along the same vein established a growing trend within the artistic community to append "in the style of X" to prompts, where X would be an artist or art movement that CLIP would ideally have knowledge of. In the following experiments, we do a systematic undertaking of prompt engineering to evaluate this family of prompts.

3 EXPERIMENT 1. PROMPT PERMUTATIONS

Before we tested any prompt, we wanted to know how many different permutations of that we would need to try to get a sense of what a prompt would return. For example, would prompting the model with "a woman in a Futurism style" lead to a significantly different generation than "a woman in a Futurist style", "a Futurist style woman", or "a woman. Futurism style"? We wanted to know: **do different rephrasings of a prompt using the same keywords yield significantly different generations?**

Our original hypothesis was that there would be no prompt permutation that would do significantly better or worse than the rest, because none of the rephrasings seemed to have significantly more meaning than the next.

3.1 Methodology

To study prompt permutations, we generated 12x12x9 images from VQGAN+CLIP (pretrained on Imagenet with the 16384 codebook) according to a combination of subjects, styles, and prompt permutations. We chose the following subjects: *love, hate, happiness, sadness, man, woman, tree, river, dog, cat, ocean, and forest*. These subjects were chosen for their universality across media and across cultures. These subjects additionally were balanced for how abstract or concrete they were as a concept as well as for positive and negative sentiment. We decided on whether a subject fell into the abstract or concrete category based upon ratings taken from a dataset of concreteness values. [7] Our set of abstract subjects averaged 2.12 on a scale from one to five (one being most abstract), and our set of concrete subjects averaged 4.80. Likewise, we chose 12 styles spanning different time periods, cultural traditions, and aesthetics: *Cubist, Islamic geometric art, Surrealism, action painting, Ukiyo-e, ancient Egyptian art, High Renaissance, Impressionism, cyberpunk, unreal engine, Disney, VSCO*. These styles likewise varied in whether they represented the world in an abstract or figurative manner. Specifically, we chose four abstract styles, four figurative styles, and four aesthetics related to the digital age. We balanced for time periods (with 6 styles predating the 20th century, and 6 styles from the 20th and 21st century).

We used these subject and style combinations to study the effect of prompt permutations: how different rephrasings of the same keywords effect the image generation.

We tested 9 permutations derived from the CLIP code repository and discussion within the online community (and give a specific example of each):

- **A MEDIUM of SUBJECT in the STYLE style** – a painting of love in the abstract style
- **A STYLE MEDIUM of a SUBJECT** – an abstract painting of love
- **SUBJECT STYLE** – love abstract art
- **SUBJECT. STYLE** – love.abstract art
- **SUBJECT in the style of STYLE** – love in the style of abstract art
- **SUBJECT in the STYLE style –love painted in the abstract art style**
- **SUBJECT VERB in the STYLE style** –love painted in the abstract style
- **SUBJECT made/done/verb in the STYLE art style** –love done in the abstract art style
- **SUBJECT with a STYLE style.** –love with an abstract art style

3.2 Annotation Methodology

We had each combination of subject and style rated by two people with backgrounds in media arts and art practice respectively. The 144 combinations were presented in 3x3 grids, where the prompt permutations were randomly arranged to prevent any effect from ordering. One combination was taken out owing to inappropriate content. Annotators were asked to note which images in the grid were either significantly better generations or significantly worse generations. We explained that they did not have to judge whether or not the generation represented the subject or the style; they just had to report whether there were generations that were significantly different from the rest. All annotators were compensated \$20/hour for however long it took them to complete the task.

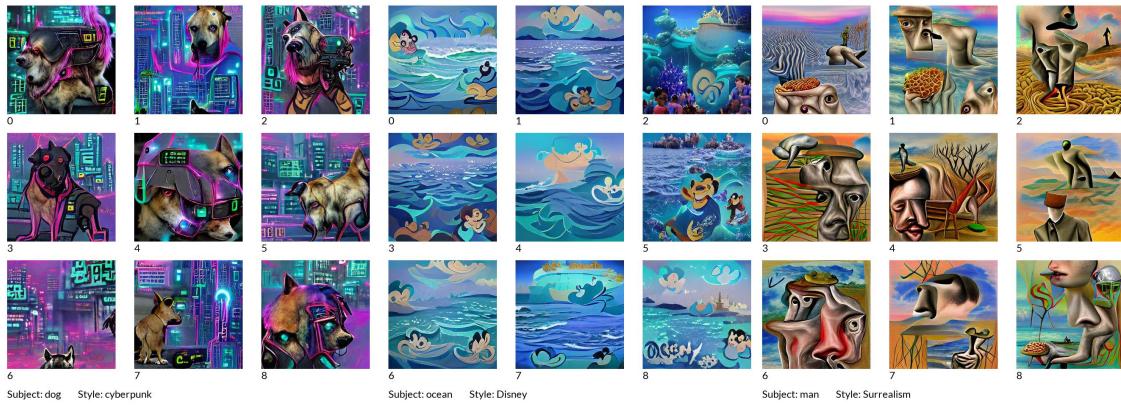


Fig. 2. For Experiment #1, annotators judged 3x3 grids where generations from different prompt permutations were arranged randomly. Annotators evaluated 143 grids of generations for significantly better generations as well significantly worse generations.

3.3 Results

We used a Fisher’s exact test to evaluate how many generations were judged as the same versus how many were judged as outliers (i.e. “significantly worse” or “significantly better”). We found that with a p-value of 0.55, the number of prompt permutations judged as outliers was insignificant when compared to the number of prompt permutations

deemed not outliers. While we report high percentage of agreement (71.3%) between our annotators, we report a low interrater reliability (Cohen’s kappa=0.0013), which we believe comes from the subjective judgment we were asking annotators to make.

We still assumed with confidence that it would be valid to test only one of the prompt permutations going forward in each of our experiments, rather than running multiple permutations of the prompt to see if the results varied significantly across permutations.

We conclude the following guideline from this experiment: **Keep the focus on keywords rather than rephrasings during prompt engineering.**

4 EXPERIMENT 2. RANDOM SEEDS

A common parameter in generative models is seeds. Generative models are stochastic, which means that it is often hard to reproduce results. To mitigate this, people often use seeds to make reproducible results. We noticed that using different seeds with VQGAN+CLIP resulted in generations were slightly different in composition. We wanted to understand: **do different seeds using the same prompt yield significantly different generations?** The motivation behind this question was to understand whether or not users would need to try exhaust multiple seeds before moving onto new combinations of keywords or know that their generations would not significantly improve from their current seed.

Our hypothesis was that no seed would do significantly better or worse than the rest, because they were random numbers.

4.1 Generation Methodology

To study the effect of seeds, we generated 1296 images from 12 subjects, 12 styles, and 9 seeds. Because neither subject nor style were the main focus of this experiment, we chose to use the same set of subjects justified in the previous section. Likewise, we chose to use the same set of styles. What we did vary was the seed chosen. We generated images using 10 randomly generated seeds 796, 324, 697, 11, 184, 982, 724, 962, and 805 and the prompt “**SUBJECT in the style of STYLE**”.

4.2 Annotation Methodology

Two annotators were shown 1296 generations in 3x3 grids where the seeds were arranged randomly. We had the nine generations varied by seed for each combination of subject and style rated by two people. Annotators were again asked to note which images in the grid were significantly better generations and significantly worse generations from the rest of the group, if any. Annotators were paid \$20/hr for as long as they needed to finish the task.

4.3 Results

We again used a Fisher’s exact test to evaluate how many generations were judged as the approximately the same versus how many were judged outliers. We found that with a p-value of <0.01, the number of generations judged as outliers was significant when compared to the number of generations deemed not outliers. Our annotators shared an inter-rater reliability of 0.13, which indicates slight agreement, which we again justify as valid given the highly subjective nature of the task (picking ‘better’ or ‘worse’ images).

This result was surprising to us, because it demonstrated that even outside of the prompt, there are stochastic components of the generation that can significantly vary the quality of the generation. We conclude from this experiment



Fig. 3. For Experiment #2, annotators judged 3x3 grids such as the ones above where generations utilizing different seeds were arranged in random order. These 143 grids were judged for significantly better generations as well as significantly worse generations.

that it is prudent to try multiple seeds during prompt engineering. A design guideline that follows from this experiment is to **try multiple generations to collect a representative idea of what prompts may return**.

5 EXPERIMENT 3. LENGTH OF OPTIMIZATION

A free parameter during each run of text-to-image models is the length of optimization: the number of iterations the networks are run for. Typically, we can expect that the more iterations, the lower and more stable the loss, and ideally the better the image. We wanted to investigate on average how many iterations are needed to get a decent result. We also wanted to see if lower iterations could be just as well generated as higher iterations, because a lower number of iterations means faster results, and for future systems involving text-to-image generation, we would want to know an average number of iterations needed to arrive at a favorable result. Our specific research question was: **does the length of optimization correlate with better evaluated generations?**

Our hypothesis was that there would be no significant difference amongst the iteration steps (100-1000), and that preferences for a preferred iteration step would be all over the map, because picking a favorite is highly subjective.

5.1 Generation Methodology

To investigate this, we tested 6 subjects (*happiness, sadness, man, cat, ocean, and forest*) across 12 styles, with a constant seed and one variety of prompt permutation. We ran the generations for 1000 iterations, and had users evaluate the generations every 100 iterations. We chose 1000 iterations as the maximum because we wanted to try a number of short to moderate wait times and 1000 is a suggested default.

5.2 Annotation Methodology

We had annotators annotate rows of generations saved at different steps of the iteration. These were specifically steps that were multiples of 100 up to 1000. The 0th iteration was not shown because the generation always began from random noise. Annotators annotated 72 rows for which generation they liked best from the set of 10. All annotators were paid \$20/hr for as long as the task took them.



Fig. 4. For Experiment #3, annotators were shown rows such as this one. These images represented different iteration steps of the optimization process. They annotated for the iteration step that they most preferred from these sets of 10.

5.3 Results

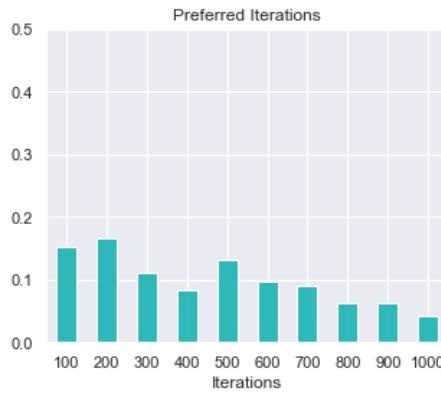


Fig. 5. Above is a plot for Experiment 3 of the frequency of times annotators chose an iteration step as their favorite iteration. Lower values of iterations tended to be preferred, and this difference over higher values of iteration was significant.

We found that the difference between the chosen iteration steps was significant upon performing a chi-squared test ($p\text{-value}=0.01$). We include the observed frequencies for the preferred iteration steps in Figure 5, where we can see that 200, 100, and 500 iterations being chosen as the most preferred.

Aggregating over the preferred frequencies, we saw that the median amount of preferred iterations was 400. We report a Cohen's kappa score of 0.33, which represents fair agreement, which we think is valid considering the highly subjective nature of picking a preferred iteration step (but more conducive to agreement than picking intuitive favorites as per the Seed experiment).

This demonstrates that a higher number of iterations did not correlate with a more desirable generation, as what might have been expected considering that more iterations optimize the multimodal loss function bringing the image and text representation closer towards one another. This is a nonintuitive result meaning that the current multimodal methods are not necessarily optimizing for generations that we prefer. Possible explanations could include the fact that lower iterations tended to be have a softer quality compared to higher iterations, where the differences and contrast seemed to be more exaggerated in higher iterations.

We conclude from this experiment **it is valid to choose lower iterations (i.e. 400 iterations) for the purposes of fast iteration.**

Medium	West Fig. Pre.	West Fig. Mod.	Non-West Abs. Mod.	Non-West Abs. Mod.	West Abs. Mod.	Internet Aesthetics
painting	Baroque	Pop Art	Ukiyo-e	Mola art	action painting	fractal
photo	High Renaissance	Surrealism	Chinese ink wash painting	Geometric Islamic art	Op art	VSCO
sketch	Impressionism	documentary photography	Kerala mural	Mexican Otomi	Bauhaus	unreal engine
cartoon	Medieval	Art deco	Mayan art	Andean textile	Cubism	ASCII art
icon	Pointillism	Hippie movement	African masks	Aboriginal art	Dadaism	Disney
vector	Neoclassicism	photorealism	ancient Egyptian art		Futurism	Studio Ghibli
graffiti			thangka			glitch
3D render						Cottagecore
						Dark academia
						Cyberpunk
						Pixar
						Pokemon

Table 1. In Experiment 4, we run 51 styles which are listed in the table above. They spanned different categories. Eight were general mediums of visual art. Twelve were Internet aesthetics. The remaining 33 were styles that were balanced for representation on both sides of the following partitions: abstract-figurative, Western and non-Western, premodern and modern.

6 EXPERIMENT 4. TESTING A BREADTH OF STYLES

In this experiment, we wanted to understand: **does the model perform equally well across a breadth of styles?**

We understood that style was a keyword we could use to suggest an aesthetic. However, we did not understand the distribution of knowledge it had learned, and we wanted to know if it would perform differently for different classes of styles.

We had a number of hypotheses for three different partitions of our set of styles.

- For abstract versus figurative styles, we assumed abstract styles would perform better because we thought they would be more tolerant to the deconstructed, global incoherence of many generations.
- For Western versus non-Western styles, we assumed that the model would perform better on Western art, since many of the computer science datasets relevant to art focus on Western schools of art (i.e. Wikiart and MetFaces).
- For styles partitioned by digital, modern, and premodern time periods, we thought digital styles would do better, as the model we used was trained on images and text from the Internet.

6.1 Generation Methodology

One of the primary goals of the paper was to investigate how comprehensively the model understood style so we could see if we could use it to structure prompts for text-to-image generation. To do so, we needed to run a battery of generations using a large number of styles.

From the perspective of art history, style represents a distinctive way in which visual arts can be grouped. To operate on this definition methodologically, we pulled styles from existing knowledge bases of art history and aesthetics online. We looked in particular at The Metropolitan Museum of Art Heilbrunn Timeline of Art History, the Aesthetics Wikia, the schema by which the WikiArt dataset was organized, and relevant Wikipedia articles to produce the set of 51 styles enumerated in the table below. These styles were chosen to balance certain factors that influence style such as time periods, culture, and whether they were abstract or figurative in the way the style represented the real world.

6.2 Annotation Methodology

Two annotators with backgrounds in media arts and art practice each received a set of subject and style combinations ordered randomly. They additionally received a directory of links to Google Images page relevant to each style. They annotated each generated combination (which was just a single image) as per the following rubric:

- 1: Extremely poor representation of the style, no motifs were present
- 2: Bad representation of the style, few motifs were present
- 3: Average representation of the style, some motifs were present
- 4: Good representation of the style, high number of motifs were present
- 5: Excellent representation of the style, very high number of motifs were present

Each annotator was instructed to judge how well the style was represented, irrespective of how well the subject turned out in the generation. Annotators were paid \$20/hr for their efforts for however long the task took them.

6.3 Results

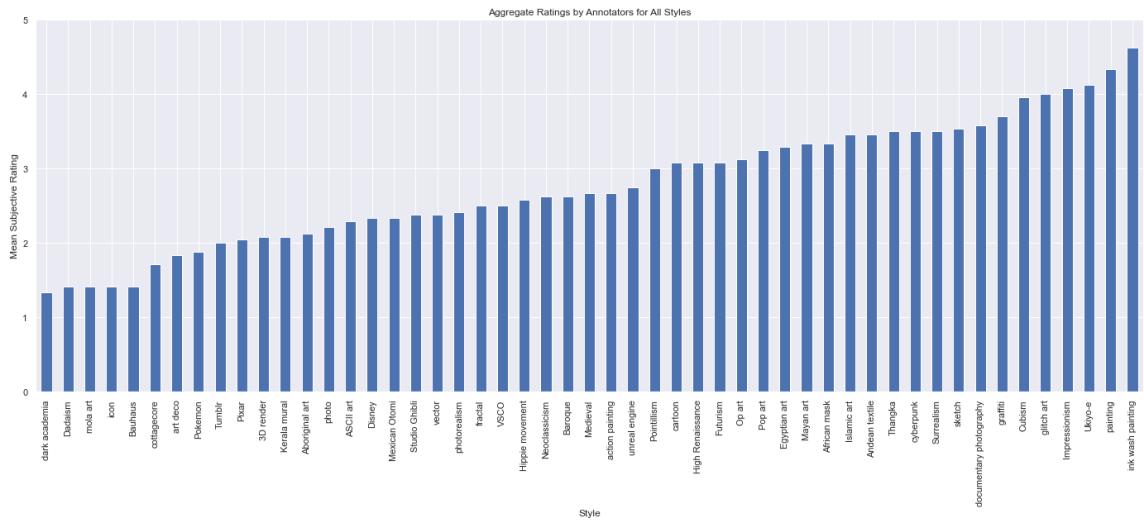


Fig. 6. For Experiment 4, 51 styles were tested across 12 subjects. This plot aggregated across subjects and ranked styles from low to high for mean subjective rating.

The model performed better on some styles than others, and we explore this variability in performance along different partitions of the data: abstract vs. figurative, Western vs. non-Western, modern vs premodern vs. digital. Figure 6 shows average ratings for all 51 styles. We first elaborate on the success modes and failure modes across all the styles as a whole before approaching the partition experiments in depth. Refer to Figure 7 for the visuals described in each of the modes of success.

6.4 Success and Failure Modes for Styles

6.4.1 Success Mode. Salient color palettes and relevant textures. The first recurring theme across successful styles was the presence of salient color scheme. This was apparent in some of the most positively judged styles such as *Ukiyo-e*, *glitch art*, *cyberpunk*, and *thangka*. These generations demonstrate that text-to-image models can match what sorts of palettes attend to what kinds of styles without being explicitly given color detail. *Cyberpunk* had a consistent global aesthetic dominated by halogen colors such as cyan and magenta, *Glitch art* pulled together colors reminiscent of TV static. Likewise, in generations such as "tree in a thangka style" or "man in the pop art style", we see different but correct understandings of the way primary colors can be saturated, contrasted, and complemented.

Design Guidelines for Text-to-Image Prompting

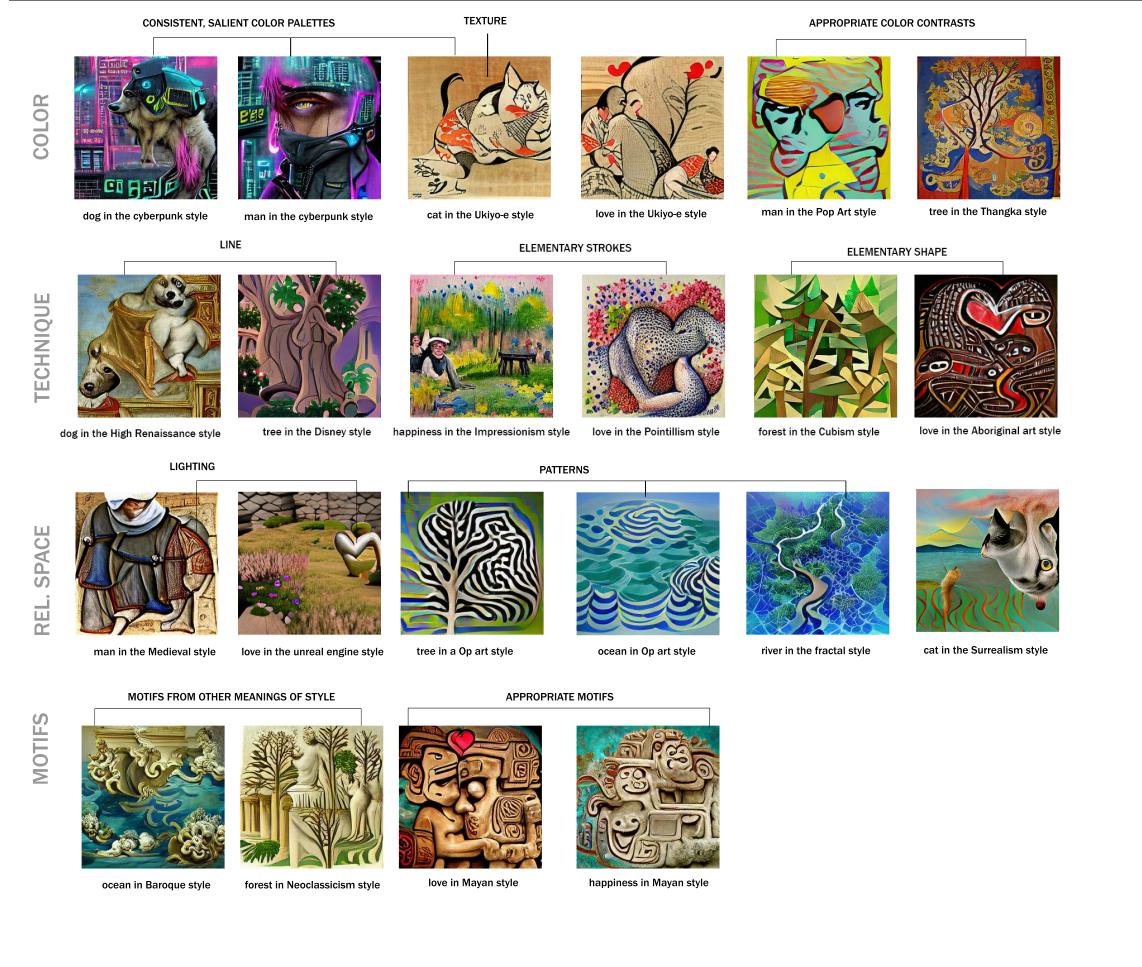


Fig. 7. In Experiment 4, we delineate different modes of success for styles that performed well in the annotation task. We see: *color*, *technique*, *relationships in space*, and *motifs*.

Another element that came through across many styles was texture. The most successful style seen from the annotations was *ink wash painting*. All generations of ink wash painting were done in wide swathes of ink that captured the watercolor quality of ink on paper. In many premodern styles such as Ukiyo-e, ancient Egyptian, Medieval, the textures of aged paper and papyrus backgrounding the image as hints of canvas also helped express the style.

6.4.2 Success Mode. Technique. Another theme across successful styles was the emulation of correct technique. Many generations exhibited choices of *line*, *texture*, and *elementary brush strokes* congruent with their style.

Across generations of the same style, the model showed the ability to use correct and consistent choices of lines. For styles such as sketch, the model produced thin lines suggestive of pencil, while for styles such as Disney or Pokemon, the model consistently produced the thick black outlines characteristic of cartoons. These lines hardly appear at all in styles such as Impressionism, which were composed instead of a patchwork of small and short strokes reminiscent of the style's broken-color technique.

In Impressionism as well as other styles like Pointillism and Cubism, the model showed that its ability to find the right style primitives. Pointillism was composed of dots and points and Cubism of deconstructed shapes. In certain styles, however, such as aboriginal art which often uses dots as their primitives, the model was able to generate textured patterns only *suggestive* of the dot primitives.

6.4.3 Success Mode. Relationships of Space. The model was always able to capture the right perspective: whether the image was done in two dimensions or three and how light and shadow were represented in the image. For example, in the Medieval style, light and shadow tended to come across flatly, while on the other extreme in styles such as unreal engine, the 3D scene lighting was very apparent in the pronounced light glares and reflections off certain elements like metal or hair.

Another component of relationships in space is composition. Styles where patterns and deconstructed gestalts were common were rated favorably. For example, the alternating patterns of swirling and swelling Black and White strips canonical to Op art were present in all Op art generations. Other examples of this success mode include the multiplicity of shapes in Cubism, recursive details in fractals, and concentric variegation in aboriginal art. This success mode could potentially have been influenced by the convolutional backbone of CLIP's image encoder, which optimizes for motifs across local neighborhoods in a way that lends well to repetition and pattern.

Figurative styles with a high tolerance for deconstructed objects such as Surrealism also performed well in the ranked annotation.

6.4.4 Success Mode. Motifs of the style. Certain styles placed motifs, or distinctive details, within the generation that made the style come through irrespective of the subject. This was especially true in styles such as the Baroque style, for which the model constantly incorporated lavish details such as ornate swirls and heavy moldings, or the Neoclassicism style generations, for which the model generated Grecian pillars and drapery in the shapes and contours of the image. It is interesting to note that while these motifs are relevant to the styles, they borrow from different facets of the meanings for Baroque and Neoclassicism that do not necessarily represent the visual arts version of the style that we had intended. We explore the ways the model misunderstood styles in the next section on failure modes.

6.5 Failure Modes

6.5.1 Failure mode: Style misunderstandings due to the multiplicity of meanings in text. As mentioned in the previous section, the model interpreted Baroque and Neoclassicist styles through the lens of decor, architecture, and sculpture, generating motifs from Baroque furniture and Neoclassical architecture and sculpture as opposed to Baroque and Neoclassical painting.

Many of the styles that performed the most poorly from the annotations were misunderstood by VQGAN+CLIP in some dimension. For example, dark academia, a social media aesthetic captured by a romanticized, Gothic approach to esoteric motifs often returned generations that contained components of a cartoon character under dramatic lighting. One possible explanation is that the model was influenced by another popular entity on the Internet that also involved

Design Guidelines for Text-to-Image Prompting

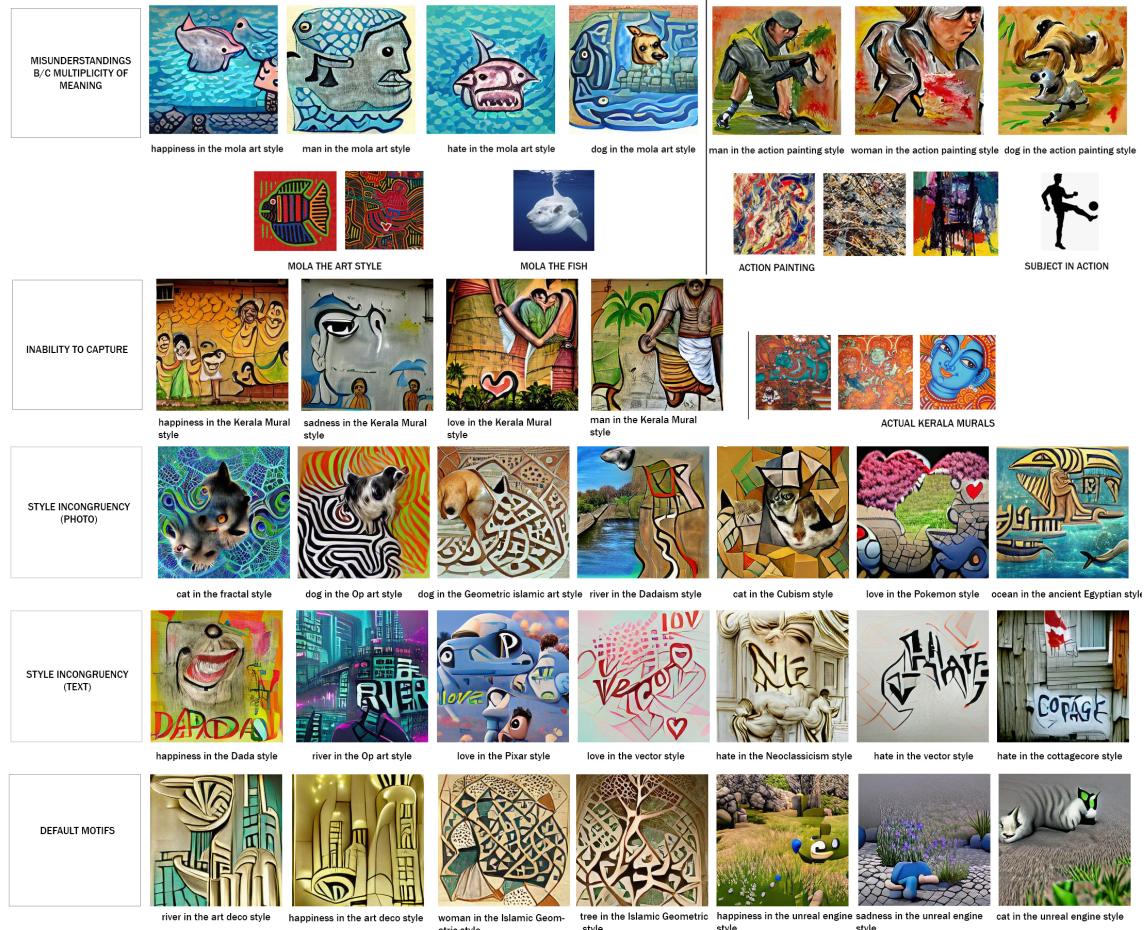


Fig. 8. For Experiment 4, we illustrate failure modes for styles such as misunderstandings owing to the multiplexity of meaning in text, inability to capture the style, style incongruencies, defaulting to motifs.

the word 'academia'—the anime My Hero Academia (the characters emerging in many of those generations shared a distinct green hair color).

Another case was the style of mola, a Latin American folk art form from Latin America with a vividly saturated color palette and heavily stylized characterization of subjects. Mola the art style was misinterpreted as mola, a species of fish. All generations of mola had a predominantly blue color palette that evoked something aquatic, and many of the subjects were blended to look like fish ("see man in the mola style" and "a man in the mola style" in Figure 8). A potential cause for understanding mola as a fish species first and foremost could be attributed to the bias towards animal species from ImageNet1000, a significant subset of which were animal species. However, the mola that were represented also were not photorealistic but rendered in a stylized form. The generations for mola as well as dark academia capture how VQGAN+CLIP generations are influenced by multiple interpretations of the style keywords.

..

Misunderstandings could also arise from multiple parsings of the prompt. For example, take the action painting art style, which is meant to refer to artists who painted dynamically using random drips and splatters. When the model generated for "a man" or "woman" in the action painting style, it created a generation that implied a *man in action* or a *woman in action*.

The many different instantiations of this failure mode suggests that the multiplicity of meanings within language both at the word level and the sentence level can present as a problem for text-to-image generation. It also represents a fundamental shift in the thought process behind the creation of a visual artwork. Visual artwork usually involves thinking about the spatial specifics of the composition, which text-to-image generation does not lend well to. Instead, text-to-image generation further problematizes the generation process with the multiple layers of meaning embedded in language.

6.5.2 Failure mode. Inability to capture styles in a complete sense. Another theme within unsuccessful styles was the inability to express styles that were more symbolic than visual. For example, Dadaism, a style representing the rejection of capitalism and embrace of the avant-garde and nonsense, was rated very poorly with a mean subjective rating of 1.42. Dadaism traditionally was expressed through satire and collaging, in manners that involved more cultural knowledge and nuanced symbolism relating to pop culture and politics.

Likewise, Bauhaus was another abstract style that was heavily influenced by abstract values rather than visual abstraction. While Bauhaus has a characteristic visual style rooted in geometric shapes, much of that is illustrated in architecture as opposed to image and in conjunction with values like harmony and utility.

These styles and their poor performance in the annotation study illustrate there are still abstractions and pools of cultural knowledge that are either not well understood or visually represented to text-to-image-models, potentially because of their different angle of abstractness. (An example of "happiness in the Dadaism style" is shown in Figure 8.)

Other styles were simply insufficiently captured. For example, if we refer to any of the images of Kerala mural style generations in Figure 8, we see that they never reached the vividly saturated and stylized look of actual Kerala mural frescos. However, they approached it, evoking color combinations and motifs such as dress that established an association of Kerala murals with India.

6.5.3 Failure mode. Style incongruity, often in the form of emerging photos or text. Sometimes elements that interrupted the style would come through the generation. These elements could be bucketed into two cases: photorealistic elements or text elements.

For example, in Figure 5 we can see a cat done in the style of Cubism. The cat's fur is entirely photorealistic, which is out of place in an otherwise abstract image. Likewise, in images of river, a photorealistic texture of a river surface would often surface even if the style was intended to be a sketch. This phenomena tended to occur for concrete subjects such as dogs, cats, rivers, and oceans.

The second case was when text began to emerge within images across iterations. For example in the generation "happiness in the Dadaism style", we see DADA explicitly written out across the generation, potentially as a compensatory technique on the model's part. Curiously, for ASCII art, text never manifested, and each generation was composed of similarly sized blurs of alphanumerical literals.

6.5.4 Failure mode: Defaulting to motifs. The model often defaulted to certain motifs. Another failure mode expressed within styles such as unreal engine and High Renaissance was a defaulting to certain motifs. For example, for the

style unreal engine, most of the text prompts returned a scene with textured rocks and grass akin to what might be encountered in a video game setting.

There were other styles for which the default motifs could not be captured particularly well. For example, Geometric Islamic artwork, Mexican otomi artwork, and digital icons all traditionally have clear-cut boundaries and hard lines between negative and positive space, silhouette and background. The generations showed a poor performance at following the implicit rules of these styles.

6.6 Results and discussion of partitions

6.6.1 Abstract versus figurative. To investigate whether the model performed better on more abstract styles or more figurative ones, we looked at a subset of 33 specific styles, excluding styles such as more general mediums (i.e. a painting, a photo) and Internet aesthetics (i.e. dark academia). We looked at a subset because these other styles generally did not fall cleanly within the dichotomy between abstract and figurative styles.

We found that abstract styles averaged a 2.63 rating (standard error 0.06), while figurative styles averaged a 3.16 rating (standard error 0.06). After running a chi squared test on the frequencies of ratings we found the difference between these ratings was significant to a p value of < 0.01.

In Figure 10 above, we visualize the top 4 styles and worst 3 styles for abstract and figurative styles. In Figure 14 we color code the ranked styles by their abstract or figurative nature.

Our original hypothesis was that abstract styles would perform better because we thought they would be more tolerant to the deconstructed, global incoherence of many generations. We found that while our original hypothesis was correct for styles such as glitch, Cubism, and Andean textiles for reasons we expected (such as a high tolerance for deconstruction), abstract styles were prone to a wide range of failure modes. These failure modes included misunderstandings due to the multiplicity of text meanings and an inability to access higher-order cultural knowledge.

The figurative styles that performed in the top 4—Ukiyo-e, Impressionism, documentary photography, and cyberpunk—displayed a diverse range of stylistic details from line to texture to perspective. The worse performing figurative styles suffered from different modes of failure such as an inability to capture the style (Kerala mural style generations) or a defaulting to unconvincing motifs (as in the case of the generation *dog in the art deco style*).

6.6.2 Western versus non-Western. To investigate whether the model would perform better on Western or Non-Western art styles, we looked again at a subset of specific styles, excluding mediums and Internet aesthetics (as they tended to be more globalized).

We found that Western art styles averaged 2.92 (standard error: 0.07), while non-Western art styles averaged 2.95 (standard error: 0.06). Using a Mann-Whitney test, we found that there was an insignificant difference between the distribution of ratings for Western styles and the distribution for ratings for non-Western styles (p-value: 0.377).

We illustrate the top performing styles in Figure 14, where we show the ranking colored for Western for Non-Western styles.

We found that the difference between Western and Non-Western styles was actually insignificant. One straightforward reason is that the Internet scale data could have compensated for the relative obscurity of any style. The alternating and even spread of the Western and non-Western styles over the x-axis of the bar graph illustrating individual styles by ranking is also visually suggestive of the same result. It is worth nothing as a potential reason for this finding that many of the non-Western styles have been long existed as objects of fascination for the West. For example, ancient Egyptian art was well curated following the phenomena of Egyptomania.

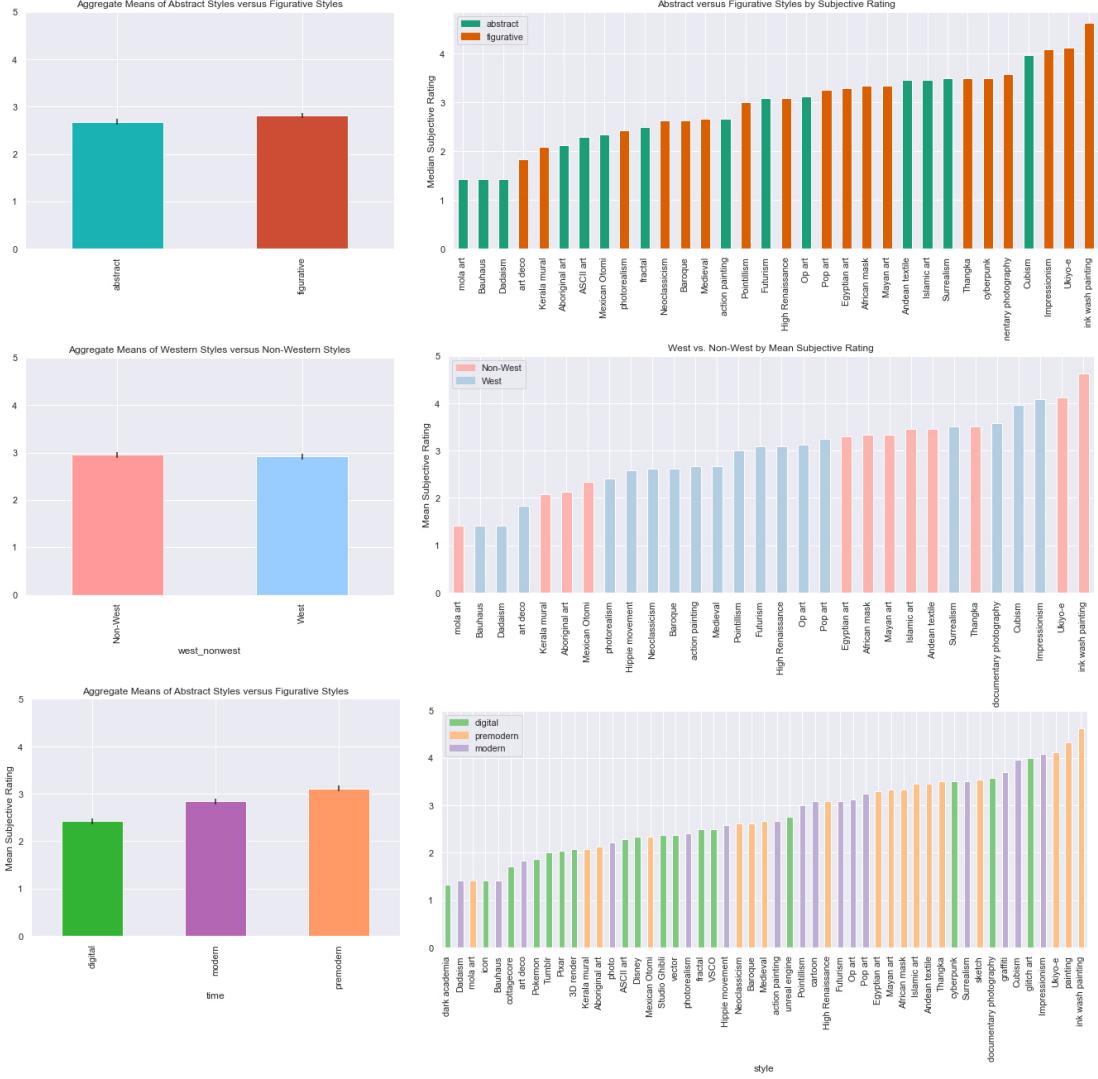


Fig. 9. For Experiment #6, in the left subgraphs, averages are reported for each category of the three partitions studied: (Abstract, Figurative), (West, Non-West), (Digital, Modern, Premodern). The right figures are bar graphs which rank each style included in the partitions by their aggregate means from low to high, coloring for their respective categories. The error bars report standard error.

6.6.3 Digital versus nondigital. We investigated whether the model would perform better on digital styles (Internet aesthetics), modern, or premodern art styles. We partitioned the styles into these time periods and colored these ranked styles by category in Figure 14.

We found that digital styles performed the worst, then modern styles, and then premodern styles with aggregate annotator ratings of 2.41, 2.83, and 3.11 respectively. Using a Kruskal Wallace test, we found these differences to be highly significant p-value < 0.001.

Design Guidelines for Text-to-Image Prompting

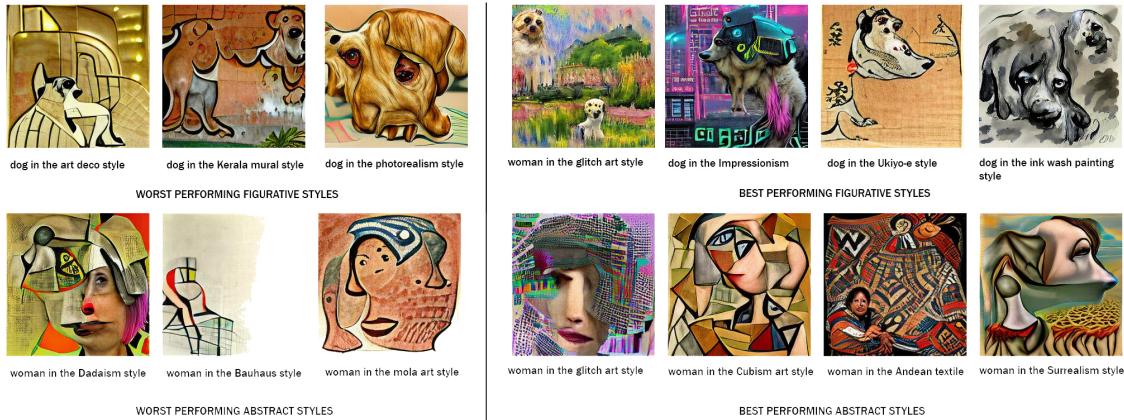


Fig. 10. For the Style Partition abstract - figurative for Experiment 4, we illustrate some of the best and worst styles.

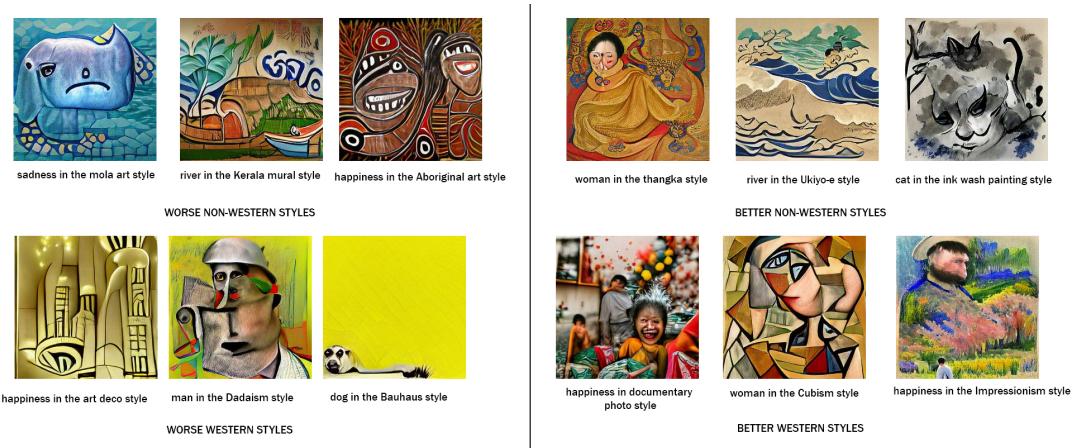


Fig. 11. For the style partition of West and Non-West, we visualized three of the best and worst styles for each group.

We found that this difference between the styles partitioned by time period was significant, and that digital styles performed the worst. One potential reason could be the digital styles had more inherent range within their styles. Some digital styles such as Tumblr, captured a photographic filter palette, while others such as *cottagecore* and *dark academia* could be represented through a number of other aesthetic forms such as fashion. Still others like Disney encompassed a range of visual styles within itself, even though the generations came across colors and lines reminiscent of the Disney Renaissance.

The Internet scale data the model was trained upon could also have retrieved a more varied set of images for the digital styles, owing to the nature of images as digital medium. Premodern and modern styles tend to exist in curated collections on digital museum archives, while digital styles exist in ubiquity across the landscape of the Internet. Likewise, styles such as Disney and dark academia could have been learned from noisier sets of images than sets of Ukiyo-e or Impressionism paintings.

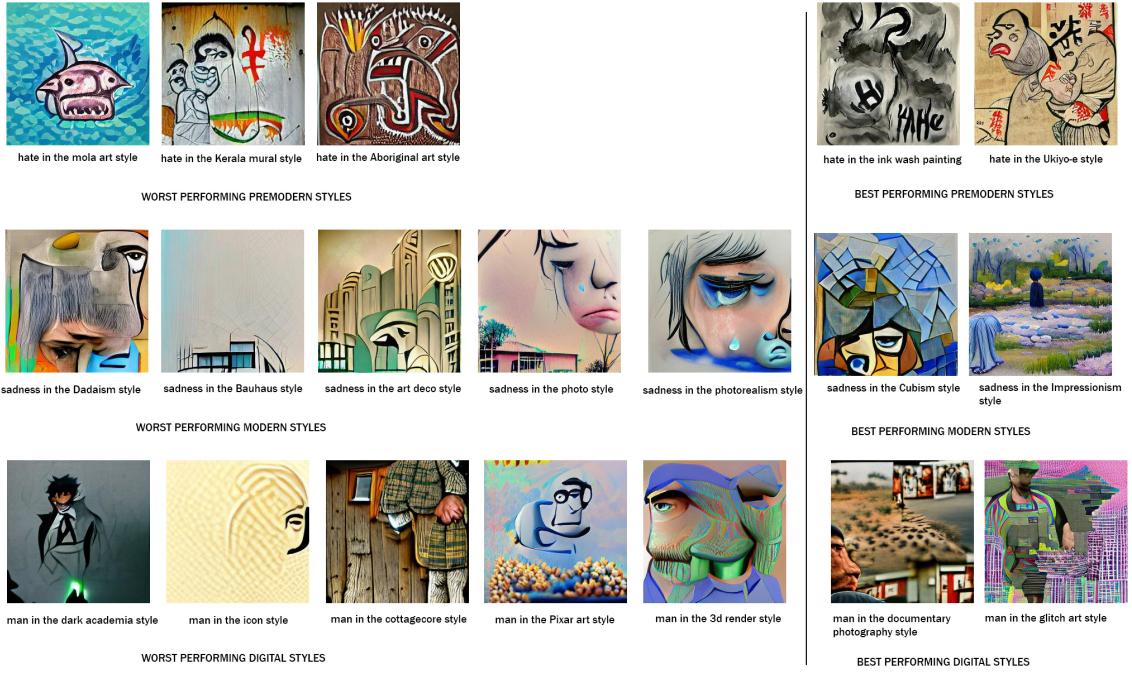


Fig. 12. For the style partition of time period in Experiment 4, we illustrate some of the best and worst styles for the categories of digital, modern, and premodern.

7 EXPERIMENT 5: INTERACTION BETWEEN SUBJECT AND STYLE

Given the success of style as a dimension for prompt engineering, we were curious to see what the interaction of subject and style was. Before undertaking this experiment, however, we ran on a pilot evaluating subject as a dimension alone. We chose not take this experiment further, because the generations yielded were too consistently poor due to the underconstrained nature of the prompt. See the appendix for further examples of this pilot.

In this experiment, we wanted to understand the following research questions: **To what degree do categories of subject and style influence one another? Do categories of styles, such as abstract or figurative styles, perform better on categories of subjects, such as abstract or concrete subjects?**

The original hypothesis we started out was that there would be a significant interaction between subject and style.

7.1 Methodology

To study the effect of interaction of subject and style, we generated 1581 images from 51 subjects, 31 styles. The full list of subjects and styles are in the appendix, but follow the same rationale as previous experiments for coverage across the abstract-concreteness spectrum (for subjects) and diversity of styles in terms of time, schools of art, and abstraction.

7.2 Annotator methodology

We recruited two annotators who had domain knowledge in art and design respectively for this task. Each received a set of subject and style combinations, ordered in different random order. They annotated each generated image for the coherency of subject and style within the image as per the following rubric:

- 1: Extremely poor representation of subject and style
- 2: Bad representation of subject and style
- 3: Average representation of subject and style
- 4: Good representation of subject and style
- 5: Excellent representation of subject and style

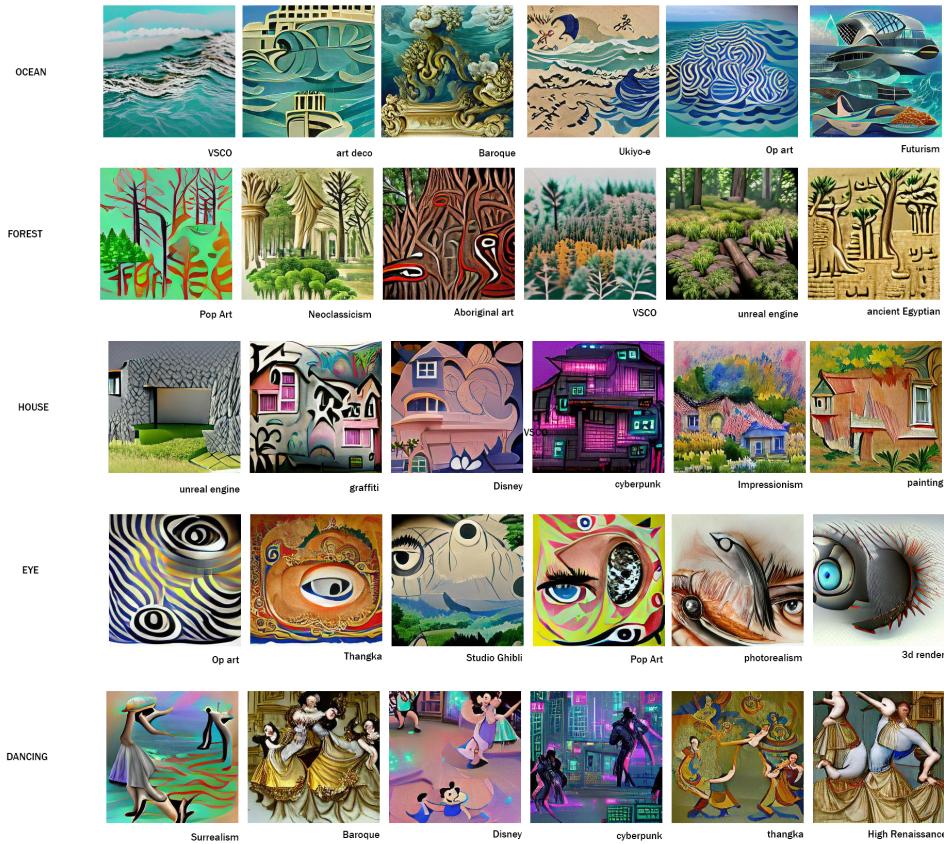


Fig. 13. For Experiment #5, generations from the top five subjects of 51 subjects are visualized above in various styles. These subjects (ocean, forest, house, eye, and dancing) were all concrete in nature.

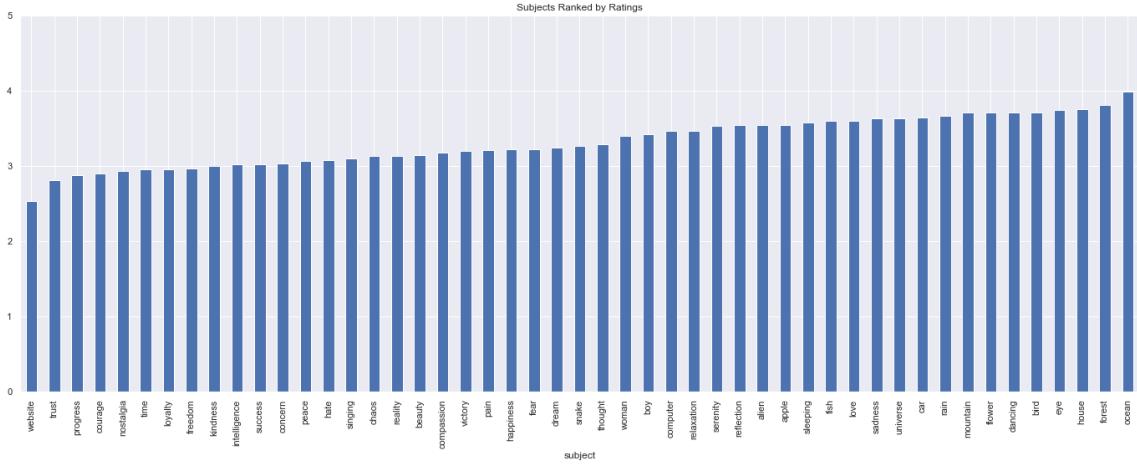


Fig. 14. For Experiment 5, 51 subjects were crossed with 31 styles. When mean rankings were aggregated across styles, the top 10 subjects all were concrete subjects. The top five specifically were ocean, forest, house, eye, and dancing. Generations of these are visualized above.

7.3 Results

Two factors we wanted to test in the experiment were the abstractness of the subject and the abstract or figurative nature of the style.

We first studied just the abstractness of the noun alone, aggregating results by subject. We found that the top ten subjects were all categorically concrete, with a concreteness value of 4.47. They were all canonical subjects within many cultures: ocean, forest, house, eye, bird. Examples of these top subjects crossed with different styles are illustrated in the image above. We found that when we compare the abstractness of the noun to the quality of the generation, there is an r value / Pearson's coefficient of 0.62, which implies a moderate to strong positive association. This means that on average there is a trend where concrete subjects tend to do better.

We then considered the influence of the abstract or figurative nature of style as well, by looking at the generations in a factorial 2x2 lens. We found the following aggregate rankings for the enumerated categories: abstract-abstract (3.05), abstract-concrete (3.17), figurative-abstract (3.49), and figurative-concrete (3.54). In running a two-way ANOVA on the annotations we found that since all p-values were significant, being well below 0.01. This allows us to conclude that both factors have a significant effect on the rating of the generation. Likewise, we saw that their interaction is also significant to a p-value well below 0.01.

8 SUCCESS AND FAILURE MODES FOR SUBJECT X STYLE

In the following section we perform a qualitative analysis on what themes in the subject-style generations produced these trends.

8.0.1 Success mode. Correct applications of symbolism. In many subjects, the text-to-image model was able to show a dexterous application of symbols. For example, in most generations for hearts, heart symbols emerged out of the image (even if symbol was incongruent with the style, for example as a heart symbol would be in Ukiyo-e art).

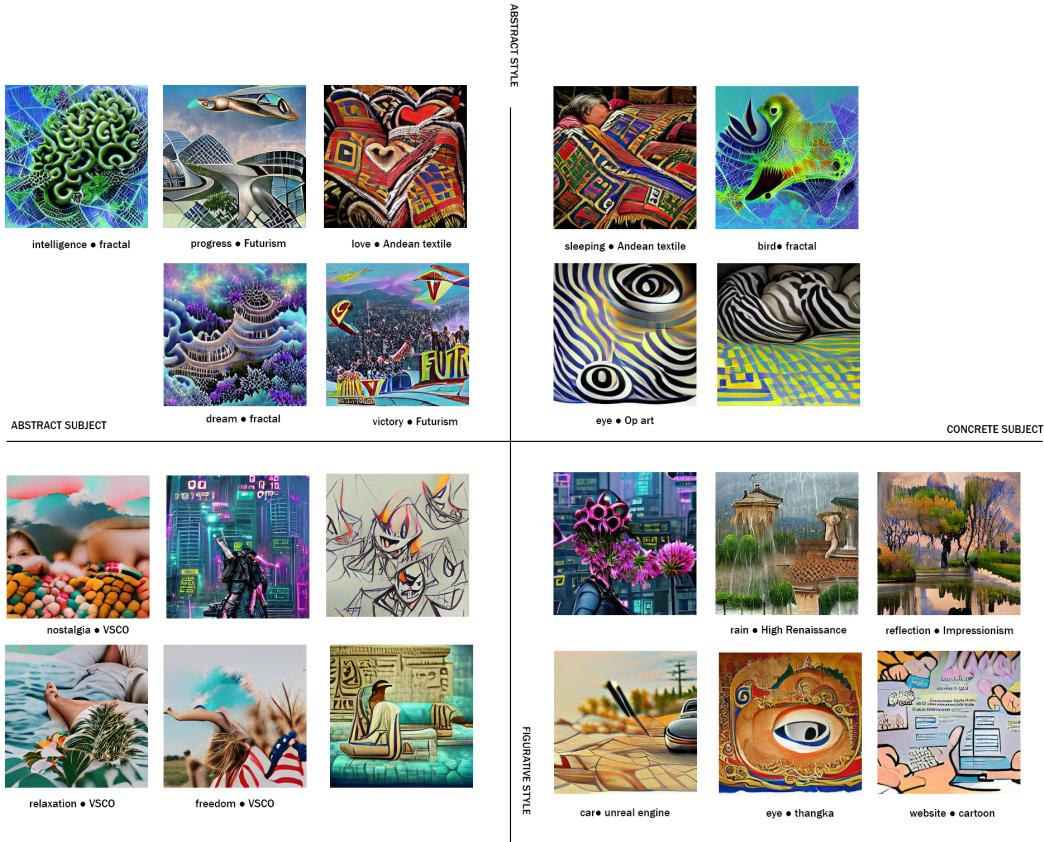


Fig. 15. In Experiment 5, we looked at the factorial interplay of subject and style. We crossed abstract and concrete subjects with abstract and figurative styles. In this figure above, we show some success cases within each crossed category.

However, generations also showed a flexible understanding of love in the form of kisses, proposals, and hugs. Generations in the subject of sadness also demonstrated an expressive range of symbols for sadness such as blueness, frowns, tears, and lonely figures. For other abstract subjects such as freedom, relaxation, or serenity, the model was able to demonstrate that it could connect freedom with American flags, relaxation with reclining. These associations are intuitive, even if certain connections, such as freedom with the United States, have overtones with implicit bias.

This success mode is primarily what makes the difference between good generations and bad generations for abstract subjects. A generation from an abstract subject is successful only when it is able to find purchase in the image as a symbol. The employment of a symbol to stand in for an abstract subject is apparent in both abstract and figurative styles.

8.0.2 Success mode. Integration of motifs with elements of the subject . Another mode of success that we could see in generations from Figure 15 such as "eye in the style of Op art" from the abstract style concrete subject category or "intelligence in the fractal style" from the abstract style abstract subject category was the blending of elementary components of subject and style. In the "intelligence in the fractal style", a symbol of a brain was chosen and rendered

with high contrast to amplify the recursive, convolutions of gray matter, thereby eliciting the sense that the brain is a fractal.

Other examples such as "flower in the cyberpunk style" or "nostalgia in the VSCO style" in the other quadrants demonstrate how the color palette of a style was colored the subject. The styles borrowed from the magenta trademark of cyberpunk and nostalgia was established through a sepia toned photo featuring pastel elements.

What makes a generation with a concrete subjects successful is when the subject is able to emerge from the abstract style without disrupting the style. For example, in the generation *rain in the High Renaissance* style in Figure 15, we see that the rain is pervasive but drawn in fine, white strokes that are characteristic to the style. Likewise, we see that the *car in the unreal engine* style applies the same effects of depth of field and scene lighting to the image prevalent in all CG renders.



Fig. 16. The subject "progress" illustrating a failure mode of the subject dropping out of certain styles (see left 3). However, nuances of progress were conveyed nonetheless in the right four images if we consider progress as progress pictures, progressivism, and something evocative of the future.



Fig. 17. Some styles exhibited the "subject dropping out" failure mode with the subject website. This subject represented a challenge because it was anachronistic to most styles.

8.0.3 Failure mode. Subject dropping out. A common failure mode, particularly for abstract subjects, was when the subject would not come through. For example, one of the most poorly rated subjects was *progress*. This is understandable, because progress is a difficult word to visualize. The images on the left side of the row illustrate nothing very relevant to progress. However, it does not mean that progress was not picked up at all by the model. The right three generations show potential interpretations of the word progress relevant to the styles of Disney, Futurism, and Hippie (through Progressive). Progress is captured in the styles as progress pictures, a potential rendition of progressivism and activism through crowds, and a vision of the future. Therefore even subjects that the model does poorly on with some styles can show nuanced understanding if there is some relevance in subject and style.

The most poorly rated of the subjects was *website*. This one is interesting because it represents a challenge to the generative model to come up with something it has likely never seen before, as computers are anachronistic to many modern and all premodern styles. We found that the model sometimes simply dropped *website* from the generation,

which we would say is neither correct nor incorrect as the subject could have conflicted with the style. This is another outcome that suggests that relevancy should be a consideration for users interacting with text-to-image generative models. However, in these generations we could also see positive outliers where the style adapted to the subject. For example, for "website in the ancient Egyptian style" seen on the in the rightmost part of the image above, we can see a keyboard, and a person using a computer screen.

8.0.4 Failure mode. Nightmare fuel and mature images. Another failure mode was when the model returned images that were either grotesque or inflammatory in nature. For example, in images using the style of High Renaissance, a common motif would be dramatic renderings of human bodies in shadow. However, because the model is in part convolutional, the motif of muscled body parts was seamlessly repeated within images until the image was cluttered with bulbous, grotesque, and interconnected limbs.

Likewise, when the authors put in for Internet aesthetics ran pilot experiments and put in prompts such as girl in the style of Tumblr or hate in the style of VSCO, the model returned images that were reminiscent of pornography and self harm. The prompts innocuously requested images for which the model added specific, unsettling details without giving any system feedback in the form of trigger warnings.

The repetitive motifs suggest that trigger warnings should be put into place for pareidolia (our subconscious tendency to read emergent forms from parts and patterns) and maturity warnings. Text-to-image generation likely translates the biases learned from the Internet into imagery.

These images are excluded from figures for the sake of propriety.

9 DISCUSSION

We condense our findings from the previous experiments into design guidelines and results to elaborate default parameters and methods for novices to interacting with text-to-image models.

- **Focus on keywords rather than the phrasings of the prompt.** Rephrasings using the same keywords do not make a significant difference on the quality of the generation as no prompt permutation consistently succeeds over the rest.
- **Try multiple generations to get a representative idea of what prompts return.** Generations may be significantly different owing to the stochastic nature of hyperparameters such as random seeds. Returning multiple results acknowledges this stochastic nature to users.
- **For fast iteration, choose lower lengths of optimization (i.e. 400 iterations).** We found that the number of iterations and length of optimization did not significantly correlate with user satisfaction of the generation.
- **Users should feel free to try a diverse set of artistic styles to manipulate the aesthetic of their generations.** The system has an impressive breadth of style information, and can be surprisingly good even for niche styles. However, avoid style keywords that have a multiplicity of meanings.
- **Pick subjects and styles that compliment each other at an elementary level.** This could be done by picking subjects with forms or subparts that could be easily interpreted by certain styles or by picking subjects that are highly relevant to the style.
- **Consider the interaction of levels of abstraction for the subject and style to may lend to incompatible representations.** Subject and style have a significant effect on one another.
- **Present users with trigger warnings for pareidolia and offensive content.** The models currently do not acknowledge the possibility for offensive content.

Our experiments gave us empirical grounding to focus many of the hyperparameters and free parameters that otherwise make text-to-image generation overwhelming, unbounded, and inexhaustive.

9.1 Implications of Borrowing Styles

Suggesting that we use pre-existing styles is at once intuitive and controversial. There are many implications to borrowing styles, one of which is that we are relying on a machine's non-expert understanding of a style to generate outputs. Sometimes the machine had a very shallow understanding of the style, as mentioned in generations for Baroque or Neoclassicism, where the style was misunderstood due to the multiplicity of meanings for Baroque in different visual mediums. Baroque paintings were known for their emphasis and dramatic chiaroscuro (application of light and shadow), but the model performed a brute force application of Baroque motifs to optimize for Baroque in the most general sense.

Another style that was shallowly summarized was Ukiyo-e. Ukiyo-e images tended to employ beige, black, and muted primary colors, however Ukiyo-e in the past was not confined to this range of color. Ukiyo-e as a style spanned centuries, during which as a style it exhibited different approaches to color ranging from monochromatic ink to brilliant brocades. Another example would be sketch- x of the y sketches were done in black and white, belying a understanding of sketches as generally black and white. This may be considered generally correct, but stereotypical.

The ability to utilize existing styles from art history and pop culture and generate stylistically relevant images represents a major achievement in visual intelligence. However, this poses significant concern for those whom those style is inherited from. Society will have to decide whether or not riffing off of certain styles is in violation of copyright and intellectual property or even potentially cultural appropriation.

One thing that is for sure is that conducting more close reads of generations and evaluations of text-to-image models can help us create methodologies to define how we identify AI-generated art from traditional art. We see a few lines of future work from here. The first is to discover learned biases that could yield variable performance across styles and subjects, in our experiments much like our own on partitions of style data.

Another direction is to work on methods that improve optimization in a way that aligns with human judgment. This takes us back to the conclusion of Experiment 3, which was that in spite of the fact that the last iteration had the lowest multimodal loss, it was not consistently picked as the favorite, meaning that there are other factors outside of the coupling between text to image that strike user fancy.

From Experiment 5, we also defined a following success mode that we saw across combinations of concrete subject, abstract styles, which was the integration of elementary elements together. Because convolutional neural networks are primed to understand low level features, we wonder if there could be utility in optimizing losses of the text and image prompt as a whole but also at lower levels of the text and image representations. That is, perhaps given the text and image encodings, "phrases" within the encodings can be optimized towards one another.

10 CONCLUSION

In this paper, we conducted a series of five experiments that each tackled a different angle of prompt engineering for text-to-image generative models involving prompt permutations, random seeds, length of optimization, style keywords, and subject and style keywords. Our experiments found significant differences between the quality of generations that fell into different categories of abstractness for style as well as subject and style. We summarized the failure and success modes of these generations through qualitative close reads of these generations. From these experiments we were

able to synthesize design guidelines to deal with the unboundedness and ambiguity of natural language as well as the stochastic nature of text-to-image interaction.

REFERENCES

- [1] 2021. <https://www.reddit.com/r/bigsleep/>
- [2] 2021. Introduction to VQGAN CLIP. <https://docs.google.com/document/d/1Lu7XPRKINhBQjcKr8k8qRzUzbBW7kzxb5Vu72GMRn2E/edit>
- [3] 2021. VQGANCLIP Hashtag. https://twitter.com/hashtag/vqganclip?src=hashtag_click
- [4] Adverb. 2021. Advadnoun. <https://twitter.com/advadnoun>
- [5] Gunjan Aggarwal and Devi Parikh. 2020. Neuro-Symbolic Generative Art: A Preliminary Study. arXiv:2007.02171 [cs.AI]
- [6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096 [cs.LG]
- [7] Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46 (10 2013). <https://doi.org/10.3758/s13428-013-0403-5>
- [8] Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. 2013. Attribit: Content Creation with Semantic Attributes. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) (*UIST '13*). Association for Computing Machinery, New York, NY, USA, 193–202. <https://doi.org/10.1145/2501988.2502008>
- [9] Katherine Crowson. 2021. Rivers Have Wings. <https://twitter.com/RiversHaveWings>
- [10] Bestiario del Hypogripho. 2021. Ayuda:Generar imágenes con VQGAN CLIP/English. https://tuscriaturas.mirahaze.org/w/index.php?title=Ayuda:Generar_im%C3%A1genes_con_VQGANCLIP/English
- [11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A Neural Algorithm of Artistic Style. arXiv:1508.06576 [cs.CV]
- [12] Songwei Ge and Devi Parikh. 2021. Visual Conceptual Blending with Large-scale Language and Vision Models. arXiv:2106.14127 [cs.CL]
- [13] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. *Mental Models of AI Agents in a Cooperative Game Setting*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376316>
- [14] Katy Ilonka Gero and Lydia B. Chilton. 2019. *Metaphoria: An Algorithmic Companion for Metaphor Creation*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300526>
- [15] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. 2019. Interactive Sketch and Fill: Multiclass Sketch-to-Image Translation. arXiv:1909.11081 [cs.CV]
- [16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]
- [17] Gwern. 2020. Gpt-3 creative fiction. <https://www.gwern.net/GPT-3>
- [18] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J. Cai. 2020. AI Song Contest: Human-AI Co-Creation in Songwriting. arXiv:2010.05388 [cs.SD]
- [19] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. *FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411764.3445093>
- [20] Justin. 2021. Somewhere Systems Twitter. <https://twitter.com/somewheresey>
- [21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakkko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. arXiv:1912.04958 [cs.CV]
- [22] Aran Komatsuzaki. 2021. When you generate images with VQGAN CLIP, the image quality dramatically improves if you add "unreal engine" to your prompt. People are now calling this "unreal engine trick" lol.e.g. "the angel of air. unreal engine" pic.twitter.com/G4xBgVlyiv. <https://twitter.com/arankomatsuzaki/status/1399471244760649729>
- [23] Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. 2019. Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum (Proc. EuroVis)* (2019).
- [24] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. *Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376739>
- [25] Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George Fitzmaurice. 2018. *Dream Lens: Exploration and Visualization of Large-Scale Generative Design Datasets*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173943>
- [26] Alexander Mordvintsev, Michael Tyka, and Christopher Olah. 2015. google/deepdream. <https://github.com/google/deepdream>
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]
- [28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs.CV]
- [29] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv:2102.07350 [cs.CL]

- [30] @someheresy Twitter. 2021. VQGAN CLIP Colab Notebook. https://colab.research.google.com/drive/1_4Jl0a7WIJeqy5LTjPjfZowMzopG5C-W?usp=sharing#scrollTo=ZdlpRFL8UAIW
- [31] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. *Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376301>

A EXPERIMENT 3

We chose to expand our list of styles to include a broader range of concreteness values, such that they included: *love, hate, peace, progress, relaxation, loyalty, compassion, beauty, pain, dream, thought, trust, freedom, chaos, success, courage, happiness, nostalgia, intelligence, kindness, time, concern, sadness, reality, serenity, fear, victory, happiness, alien, car, house, apple, singing, dancing, sleeping, mountain, rain, ocean, forest, flower fish, bird, snake, boy, woman, eye, computer, website, universe, reflection*. Likewise, we chose to use the same set of styles: *photorealism, Studio Ghibli, Neoclassicism, African mask, thangka, fractal, Hippie movement, ancient Egyptian, art deco, unreal engine, Disney, cartoon, Pop art, VSCO, Futurism, 3D render, Pointillism, sketch, Surrealism, Andean textile, Aboriginal art, Ukiyo-e, High Renaissance, Mayan, graffiti, Cubism, Impressionism, Baroque, Op art, cyberpunk, and painting*.

B EXPERIMENT 5. SUBJECTS

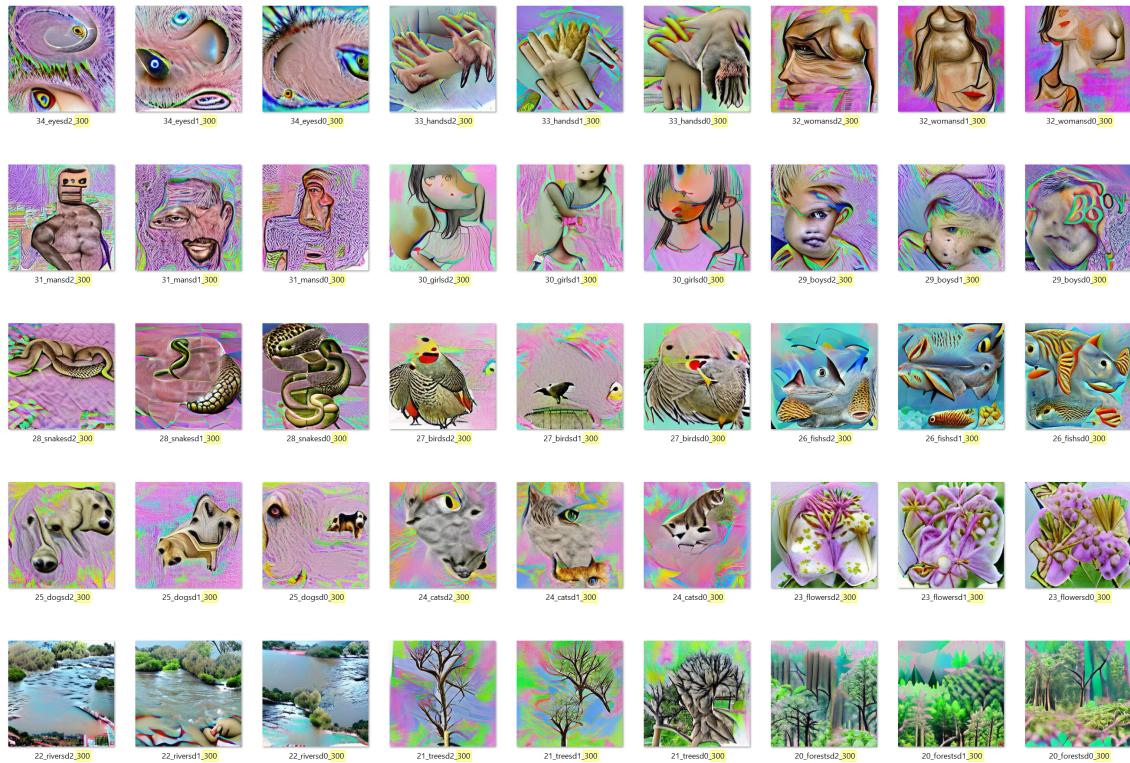


Fig. 18. In a pilot before Experiment 5, we found that using only subjects as keyword dimensions was insufficient. The underconstrained nature of the generation made generations too poor to evaluate, because they were not grounded in any sort of aesthetic.