Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering

JONAS OPPENLAENDER, University of Jyväskylä, Finland
RHEMA LINDER, University of Tennessee, United States
JOHANNA SILVENNOINEN, University of Jyväskylä, Finland

Humankind is entering a novel era of creativity — an era in which anybody can synthesize digital content. The paradigm under which this revolution takes place is prompt-based learning (or in-context learning). This paradigm has found fruitful application in text-to-image generation where it is being used to synthesize digital images from zero-shot text prompts in natural language for the purpose of creating AI art. This activity is referred to as prompt engineering — the practice of iteratively crafting prompts to generate and improve images. In this paper, we investigate prompt engineering as a novel creative skill for creating prompt-based art. In three studies with participants recruited from a crowdsourcing platform, we explore whether untrained participants could 1) recognize the quality of prompts, 2) write prompts, and 3) improve their prompts. Our results indicate that participants could assess the quality of prompts and respective images. This ability increased with the participants' experience and interest in art. Participants further were able to write prompts in rich descriptive language. However, even though participants were specifically instructed to generate artworks, participants' prompts were missing the specific vocabulary needed to apply a certain style to the generated images. Our results suggest that prompt engineering is a learned skill that requires expertise and practice. Based on our findings and experience with running our studies with participants recruited from a crowdsourcing platform, we provide ten recommendations for conducting experimental research on text-to-image generation and prompt engineering with a paid crowd. Our studies offer a deeper understanding of prompt engineering thereby opening up avenues for research on the future of prompt engineering. We conclude by speculating on four possible futures of prompt engineering.

CCS Concepts: • **Applied computing** → *Fine arts*; • **Human-centered computing** → **Interaction paradigms**.

Additional Key Words and Phrases: prompt engineering, prompting, text-to-image generation, AI art, creativity

1 INTRODUCTION

We are entering an era in which anybody can generate digital images from text — a democratization of art and creative production. In this novel creative era, humans work on "prompt-based engineering" within a human-computer co-creative framework [25]. Emerging digital technologies will co-evolve with humans in this digital revolution, which requires the renewal of human capabilities and competences [16]. One increasingly important human skill is "prompting" due to it providing an intuitive interface to AI. Prompting (or "prompt engineering") is the skill and practice of writing inputs ("prompts") for generative models [33, 38]. Prompt engineering is iterative and interactive — a dialogue between humans and artificial intelligence (AI) in an act of co-creation. As generative models become more widespread, prompt engineering has become an important research area on how humans interface with AI [3, 8, 9, 19, 21, 33, 48].

One area where prompt engineering has been particularly useful is the field of digital visual art. State-of-the-art image generation systems, such as OpenAI's DALL-E [46], Midjourney [36], and Google's Imagen [50], have been trained on large collections of text and images collected from the World Wide Web. These systems can synthesize high-quality images in a wide range of artistic styles [12, 33] from textual input prompts. Practitioners of text-to-image generation often use prompt engineering to improve the quality of their digital artworks [33]. Within the community of practitioners, certain keywords and phrases have been identified that act as "prompt modifiers" [39] [39]. These keywords can, if included in a prompt, improve the quality of the generative model's output or make images appear in a specific artistic style [12, 33, 38]. While a short prompt may

already produce impressive results with these generative systems, the use of prompt modifiers can help practitioners unlock the systems' full potential [38, 39, 56]. The skillful application of prompt modifiers can distinguish expert practitioners of text-to-image generation from novices.

Whether prompt engineering is an intuitive skill or whether there is a learning curve to it has, so far, not been investigated. There are a number of reasons why such an investigation is important. A look at StableDiffusion's Discord channel[1] shows preliminary evidence that some prompts and keywords combinations circulating in the community of practitioners do not seem intuitive to novice users. Such keywords include, for instance, the modifier *"by Greg Rutkowski"* [39] or other popular modifiers, such as *"smooth," "elegant," "luxury," "octane render,"* and *"artstation"* to boost the quality of an image [39]. These modifiers are intuitively applied by practitioners in the AI art community, but may confront novices with challenges of understanding their effect on the resulting image. A further source for unintuitiveness of the current practice of prompt engineering is a misalignment between the written human prompt and the way in which the text-to-image models interpret the prompt. Compared to how we humans understand a prompt and its constituents, the text-to-image model may attach very different meanings to some keywords in the prompt. The CLIP Interrogator[2] image captioning tool gives us a glimpse into what a text-to-image model "sees" in an image. Using this tool on any given image will result in unintuitive keyword combinations that a human user would likely never have chosen. Further confounding the problem is that keywords in prompts can affect both the subject and style of a generated image.

Whether prompt engineering is a skill that humans apply intuitively or whether it is a learned skill is important not only for the field of AI art, but also for research on human-AI interaction and the future of work in general. Today, many images are shared on social media, often with stunning results. If what we see on social media is the result of the application of prompt engineering by experts, then the generative content that we encounter on social media could be skewed by a small group of highly skilled practitioners. From a systemic perspective, we run the risk of assigning too much importance to prompting as a method for interacting with generative models if prompt engineering is an intuitive skill [41]. On the other hand, if prompt engineering is a learned skill that requires expertise and training, this could give rise to novel creative professions with implications for the future of work.

In this paper, we explore the creative skill of prompt engineering in three studies with a total of 227 participants recruited from Amazon Mechanical Turk (MTurk), a popular microtask crowdsourcing platform. The studies are summarized in Table 1.

Table 1. Overview of the three studies.

Study No.	No. of participants	Study purpose	Research question
1	52	Test participants' understanding of prompt quality	Are participants able to tell the quality of an image from the textual input prompt?
2	125	Test participants' ability to write prompts	Can participants effectively write prompts to create digital artworks?
3	50	Test participants' ability to revise their own prompts	Can participants improve their prompts to generate better digital artworks?

[1] https://discord.com/channels/1002292111942635562/
[2] https://github.com/pharmapsychotic/clip-interrogator

In Study 1, we explore participants' understanding of how a text-to-image system produces images of varying quality depending on the phrasing of input prompts. A feeling of what contributes to the quality of a prompt could enable participants to write prompts and create high-quality images on their own. In our within-subject experiment, participants separately rated the aesthetic appeal of textual prompts and matching images generated with a text-to-image system. We hypothesize that a high degree of consistency within the participants' two ratings may point toward there being a strong understanding of what makes a "good" prompt.

In Study 2, we invite participants to put their knowledge and expertise into practice by writing three input prompts for a text-to-image system with the specific aim of creating a digital artwork. We analyze participants' use of descriptive language and the use of prompt modifiers that could influence the quality and style of the resulting artworks. In Study 3, we then invite the same participants who participated in the previous study to review the images generated from their own prompts. Each participant improved the prompts with a specific task of creating an artwork of high visual quality. With this study, we investigate whether expertise in writing prompts emerges intuitively or whether it is an expert skill, learned through iteration and practice. Our hypothesis is that if prompt engineering is a learned skill, participants will not be able to significantly improve their images due to few interactions with the text-to-image system within our studies.

We find that while participants were able to describe artworks in rich descriptive language, almost none of the participants used specific keywords to adapt the style of their artworks or modify the images in other ways. Moreover, participants were not able to significantly improve the quality of the artworks in the follow-up study. This points to prompt engineering being a non-intuitive skill that laypeople first need to learn before it can be applied in meaningful ways.

Due to our decision to recruit crowd workers as participants, our paper is the first to provide insights on how well paid crowd workers perform in experiments on prompt engineering and text-to-image generation. Based on our findings and experience, we provide recommendations for conducting experiments on text-to-image generation and prompt engineering with an extrinsically motivated crowd. We conclude by speculating on four potential futures for prompt engineering.

2 RELATED WORK

2.1 Text-to-Image Generation with Deep Learning

Text-to-image generation is a type of deep learning technology that allows users to create images from text descriptions. This technology has gained significant interest since early 2021, when OpenAI published the results of DALL-E [47] and the weights of their CLIP model [45]. CLIP is a multi-modal model trained on over 400 million text and image pairs from the Web. The model can be used in text-to-image systems to produce high-fidelity images. Many approaches and architectures for image generation with deep learning have since been developed, such as diffusion models [11]. These approaches typically use machine learning models trained with contrastive language-image techniques using training data scraped from the Web. These systems are text-conditional, meaning they use text as input for image synthesis. This input, known as "prompt," describes the image to the system, which then generates one or more images without further input.

2.2 Prompt Engineering

The practice of crafting input prompts is referred to as prompt engineering (or prompting for short). In this section, we explain the 'engineering' character of prompt engineering and highlight the difference to automated approaches of prompt optimization.

2.2.1 The engineering character of prompting. The term prompt engineering was originally coined by Gwern Branwen in the context of writing textual inputs for OpenAI's GPT-3 language model [33].

'Engineering,' in this case, does not refer to a hard science as found in science, technology, engineering, and mathematics (STEM) disciplines. Prompt engineering is a term that originates from within the online community of practitioners. Practitioners include artists and creative professionals, but also novices, amateurs, and more serious "Pro-Ams" [17] aiming, for instance, to sell their creations as digital art based on non-fungible tokens (NFTs) [28]. Not every member of this online community may identify as a prompt engineer. An alternative self-understanding could be "promptist" [22] or "AI artist" [58]. One aspect of prompt engineering that relates to its engineering character is that it often involves systematic experimentation through trial and error [33]. The challenge for the prompt engineer is not only to find the right terms to describe an intended output, but also to anticipate how other people would have described and reacted to the output on the World Wide Web.

2.2.2 Difference to soft prompting techniques. Prompt engineering is a language-based practice conducted by humans who write prompts in discrete tokens. Thus, it differs from so-called "soft prompting" approaches that aim to automatically optimize input for machine learning models. For example, prefix tuning [32] optimizes continuous vectors to fine-tune pre-trained language models for downstream tasks, but it operates in vector space with "virtual tokens" rather than discrete tokens. Likewise, prompt tuning [30] uses backpropagation to learn input prompts for a frozen language model to perform downstream tasks. Prompt optimization [10] uses reinforcement learning to optimize prompts. In contrast to the above, prompt engineering involves manually writing and rephrasing prompts. As such, prompt engineering is closer to human-centered fields, such as Human-Computer Interaction, Computer-Supported Cooperative Work, Human-AI Interaction, and conversational AI than to the field of machine learning.

2.3 Prompt Engineering for AI Art

"AI art" [58] — or art generated by artificial intelligence — has become a popular application for prompt engineering [38]. An online community has formed, sharing images and prompts on various platforms. Within this community, certain practices for writing prompts have emerged. For example, prompts often follow a specific pattern, such as the following template [52]:

> *[Medium] [Subject] [Artist(s)] [Details] [Image repository support]*

A typical prompt could be [1]:

> *A beautiful painting of a singular lighthouse, shining its light across a tumultuous sea of blood by greg rutkowski and thomas kinkade, Trending on artstation.*

Prompt modifiers, such as the underlined words above, are added to a prompt to influence the resulting image in a specific way [33, 38, 39]. Prompt modifiers are an important technique in prompt engineering for AI art because they allow the prompt engineer to control the output of the text-to-image system. Prompt modifiers may make the resulting images subjectively more aesthetic and attractive [33, 38].

Different types of prompt modifiers are used in the AI art community [39], but the two most common types of modifiers affect the style and quality of images. These prompt modifiers consist of specific keywords and phrases that have been found to modify the style or quality of an image (or both). Modifiers that affect the quality of images are known as quality boosters [39], and can include phrases such as *"trending on artstation," "unreal engine," "CGSociety," "8k," and "postprocessing."* Style modifiers affect the style of an image and can include a wide variety of open domain keywords and phrases, such as *"oil painting,", "in the style of surrealism," or "by James Gurney"* [39].

Human-centered research on prompt engineering for text-to-image synthesis is still in its early stages, with only a few papers published on the topic in the field of Human-Computer Interaction

(HCI). Liu and Chilton's study on subject and style keywords in textual input prompts mentioned that without knowledge of prompt modifiers, users must engage in "brute-force trial and error" [33]. The authors presented design guidelines to help people produce better results with text-to-image generative models. Qiao et al. conducted an experiment on using images as visual input prompts, resulting in design guidelines for improving subject representations in AI art [44]. Besides these guidelines, there are also many community-provided resources that offer guidance for novices and practitioners of AI art, such as Ethan Smith's "Traveler's Guide to the Latent Space" [52], Zippy's "Disco Diffusion Cheatsheet" [1], and Harmeet Gabha's "Disco Diffusion Artist Studies" [12]. These resources provide a wealth of information about prompt modifiers for producing high-quality visual artifacts.

3 STUDY 1: UNDERSTANDING PROMPT ENGINEERING

We conducted an experiment to study participants' understanding of prompt engineering. Participants were asked to rate both the AI-generated images and the corresponding textual prompts. We hypothesized that participants with a strong understanding of prompt engineering would exhibit a high consistency between the ratings in the two modalities. In other words, if someone can predict the aesthetic appeal of an image from its textual prompt, they likely have a good sense of how prompt engineering works. The study design reflects the knowledge that prompt engineers would use in practice. A good understanding of textual prompts is crucial for predicting how well a prompt will perform. Longer prompts that include descriptive language and many modifiers are likely to produce higher quality artworks than short ones.[3] In the following section, we describe how we selected a set of prompts and images for this study.

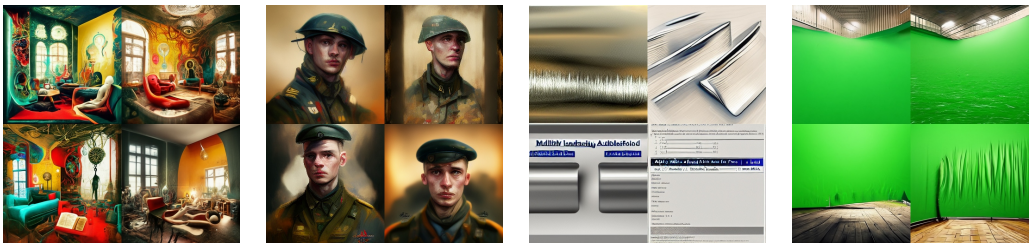
3.1 Method

3.1.1 Research Materials. We curated a set of prompts and images created with Midjourney, a text-to-image generation system and community of AI art practitioners. Using purposeful sampling, we selected 111 images from the corpus of over 3500 images generated by the first author on Midjourney. Our choice to use the author's corpus has several advantages. The corpus includes images with a range of different prompt modifiers commonly used on Midjourney and we avoid intruding on others' intellectual property rights. Further, the author has experience with text-to-image generation and can distinguish failed attempts from successful ones. This allowed us to create a corpus of images with varying levels of subjective quality. Specifically, we selected 59 images judged as failed attempts and 52 images of high aesthetic quality. We kept the format of four images per prompt, as it resembles the output a prompt engineer would typically receive on Midjourney.

To assess the aesthetic quality of the 111 images in the dataset, we recruited ten raters from two academic institutions. The raters had diverse backgrounds in Computer Science, Information Sciences, Human-Computer Interaction, Cognitive Science, Electrical Engineering, and Design. They consisted of 2 Professors, 3 PostDocs, 3 PhD students, 1 Master student, and 1 project engineer (5 men and 5 women, age range 24–48 years). Raters completed a simple binary classification task to classify the images as high or low quality based on their aesthetic appeal. Raters were informed that there was an unequal number of images in each category. The inter-rater agreement over all images, as measured by Fleiss' kappa, was fair, $\kappa = 0.34$, $z = 23.9$, $p < 0.00$, 95% CI [0.31, 0.37].

We discussed the ratings and selected images for further study. From the set of images with perfect agreement among raters, we selected ten high- and ten low-quality images. The final set contains 20 images and respective prompts of varying quality (see Figure 1 and Appendix A).

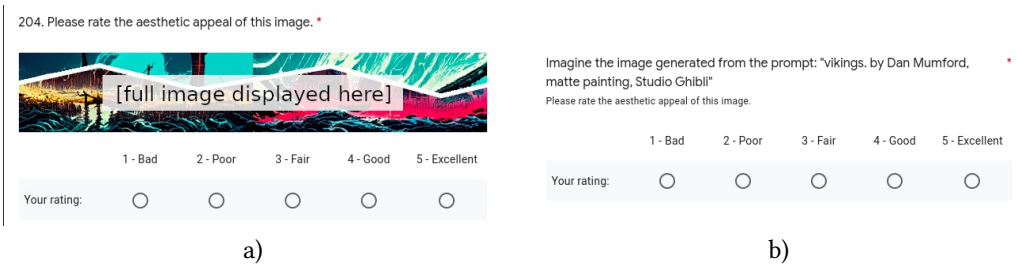
[3]Note that some state-of-the-art image generation systems, like Midjourney version 4, are "greedy" and will try to turn any input into an aesthetic artwork, even if the prompt is short or non-descriptive. See Section 6.4 for more on this issue.



1a) High aesthetic appeal 1b) Low aesthetic appeal

Fig. 1. Exemplars of images used in Study 2. The full set of images and prompts is listed in Appendix A.

3.1.2 Study Design. We conducted a within-subject experiment with two conditions. In the first condition, participants rated 20 AI-generated images (see Figure 2a) on a 5-point Absolute Category Rating (ACR) scale [43, 51]. The ACR is known to produce reliable judgments [51] that are relatively insensitive to environmental factors, such as lighting, monitor calibration, language, and country [43]. In the second condition, participants were asked to imagine the images that would result from 20 textual prompts used to generate the images in the previous condition, and rate them using the same scale. In this condition, participants were only shown the prompts, not the images. This part of the study was designed to make participants focus on the resulting images, rather than the prompts themselves. To help with this, each prompt was prefaced with "Imagine the image generated from the prompt: …" and a descriptive line was added as a reminder of the task (see Figure 2b).



a) b)

Fig. 2. Example of items used in Study 2 for a) rating the performance of prompts in generating aesthetic images and b) aesthetic appeal of AI-generated artworks.

The instructions were carefully designed to avoid confounding factors. For instance, we used neutral wording and avoided referring to the images as artworks to prevent higher positive aesthetic ratings [2, 15, 42, 54]. Our aim was not to measure exact ground-truth ratings for aesthetic appeal, but to study differences in ratings within participants.

After the two conditions, we collected basic demographics including participants' experience with art, art practice, and text-based image generation. We also included an optional open-ended item for participants to elaborate on their experience with text-to-image generation. Experience with art and text-based image generation were measured on 5-point Likert scales, and experience with art was measured as a binary variable.

3.1.3 Participant Recruitment and Procedure. We recruited US-based participants from Amazon Mechanical Turk (MTurk) with a task approval rate greater than 95% and at least 1000 completed tasks. This combination of qualification criteria is common in crowdsourcing research (e.g., Hope

et al. [18]). The experiment was implemented as a survey task and hosted on Google Forms. Participants were paid US$1.50 for completing the survey. The price was determined from the average completion times in a small-scale pilot study ($N = 9$, US$1 per task).

The task consisted of 31 items in total, including a consent form, an introduction to the study, 20 ratings of prompts, 20 ratings of images, ten demographic items, and one consistency check. Participants underwent the two conditions (rating of prompt and rating of images) in balanced order. Half of the participants first rated the prompts, then the images, and the other half vice versa.

To prevent bias, we anonymized the filenames of the images and assigned a random numeric code to each image to make it harder to associate the images with the prompts from the previous survey section. As a check for consistency, we duplicated one image and collected a rating for this image (L1, see Appendix A.2). Participants who differed in their rating by greater than one category on the ordinal ACR scale were excluded from analysis. We excluded four participants for failing this consistency check and another two participants for having completed the survey without completing the task on MTurk. The final sample included 52 participants.

3.1.4 Analysis. We compared the distributions of ratings using a Kruskal-Wallis rank sum test followed by posthoc Dunn tests where applicable. This analysis showed whether the type (Prompt or Visual Artwork) and quality (High or Low) had different ratings. Then we performed a correlation test using Pearson's product-moment correlation to look at the relationship between paired scores for each type. We hypothesize that participants can detect a relationship between the quality of a prompt, in terms of its ability to depict visual art through human imagination, and the quality of the visual artwork generated by the text-to-image system. Further, we hypothesize that art experience, as measured by self reports, will have an impact on the consistency of the participants' ratings. To test this hypothesis, we tested the correlation between art experience and average error for each participant. We calculated the participants' art experience by summing the self-reported "practice of art" (1 to 5) and "museum visitation" (1 to 5), but not text-to-image generation experience (because only one participant had strong experience, which they obtained from taking an MTurk survey). Average error per participant was calculated by taking the average absolute difference between each pair of prompt and artwork rating. For example, if all prompts were rated as 2 and all artworks as 5, the average error would be 3.

3.2 Results

3.2.1 Participants. Participants ($N = 52$) were between 24 and 67 years of age ($M = 38.2$ years, $SD = 12.98$ years) and included 31 men and 21 women (no non-binary) from diverse educational backgrounds (27 Bachelor's degrees, 10 Master's degrees, among others). A sizable fraction of the participants (46%) reported having an educational background in the arts. Twenty-nine participants agreed and nine strongly agreed that they had visited many museums and art galleries ($M = 3.60$, $SD = 1.18$). However, participants did not practice art often ($M = 3.08$, $SD = 1.28$). Overall, participants were interested in AI generated art ($M = 3.69$, $SD = 0.83$), but had little experience with text-to-image generation ($M = 2.58$, $SD = 1.43$). Only three participants mentioned having used text-to-image generation (DALL-E mini/Craiyon) before.

3.2.2 Visual and prompt ratings. Our study design asked participants to rate both Prompts and Visual artwork. As these crowd workers did not receive special training for prompt engineering or text-based AI art, our goal was to understand the quality of our participants' perceptions. We show the histogram of scores broken into groups for each Art Type (Prompt and Artwork) in Figure 3 and Table 2 and the Quality (high or low) as described previously. Visually, these show differences across groups, with the distributions of Artworks leaning towards higher quality than prompts.

Table 2. Average rating across Art Type and Quality.

Art Type	Quality	Mean	Std Dev
Artwork	High	3.70	1.04
Artwork	Low	3.39	1.15
Prompt	High	3.87	1.07
Prompt	Low	2.78	1.28

We used a Kruskal-Wallis rank sum test across these four unique groups, finding a significant difference ($\chi^2 = 231.4$, $p < 10^{15}$, $df = 3$). Following this significant result, we performed post-hoc Dunn's test pairwise across each group with Bonferroni correction for p-values. Each of these pairs had significant results with a p-value of less than 10^{-4}, except for Artwork–High versus Prompt–High, in which $p < .004$. This implies that the median values among all comparisons of groups (i.e. Artwork–High, Artwork–Low, Prompt–High, Prompt–Low) are significantly different from each other.

Participants were able to differentiate images with low visual aesthetic quality from high quality images. Artwork–High has a higher mean rating ($\mu = 3.70$) compared to Prompt–Low ($\mu = 3.39$). Likewise, participants were able to distinguish between high and low quality by imaging what would be produced based on textual prompts. Prompt–High has a higher mean rating ($\mu = 3.87$) compared to Prompt–Low ($\mu = 2.78$). The overall span between the Artwork High and Low is larger for Prompts ($3.87 - 2.78 = 1.09$) than for Artworks ($3.70 - 3.39 = .31$). Both High and Low quality Artworks had distributions that favored a rating of 4, while Prompt–Low has a relatively flat distribution across values of 1 to 4 (see Figure 3).

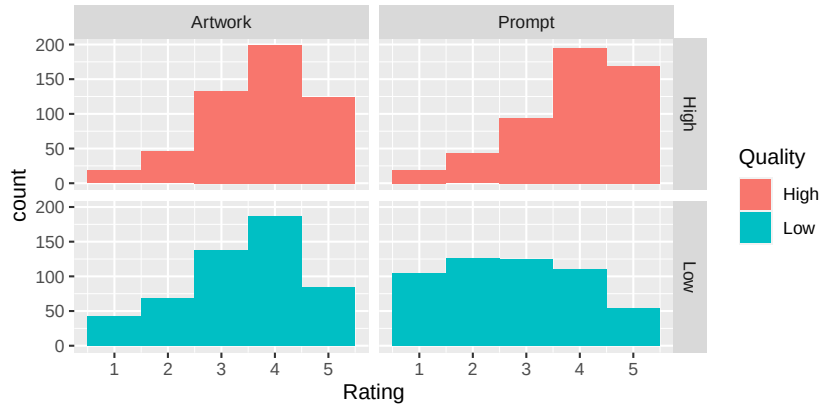


Fig. 3. Scores by participants ($N = 52$) for images (a) and prompts (b) with high aesthetic quality (H1–H10) and low aesthetic quality (L1–L10) in Study 2.

3.2.3 Connection between visual image and prompt quality. While in theory, prompts that can help readers conjure (i.e. visualize or imagine) more aesthetically appealing mental images will also generate better Artwork, it is not clear whether this would be the case for untrained crowd workers. While our participants were not able to directly associate Prompts to Artworks, each Artwork had a matching Prompt. We used a Person's product-moment correlation test to measure whether

ratings for the Prompt and Artwork are correlated. The test shows a weak (r = .29, 95% CI [.23, .34]) but significant ($p < 10^{-15}$) positive correlation between ratings from Artworks and Prompts. This indicates that when a Prompt is seen as having a higher quality, that it is also more likely that the Artwork will appear as having a high quality.

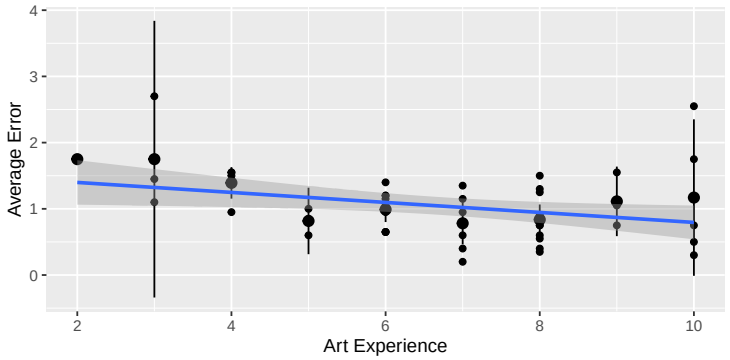


Fig. 4. As art experience increases, the average error in differences between the textual Prompt and visual Artwork decreases.

3.2.4 Art familiarity and consistency. We hypothesized that when participants are more familiar with art, they will be more effective at rating art. One way to measure this is to take the average absolute difference between each Prompt and visual Artwork pair for each participant. Being able to visualize and rate Prompts and Artworks with equal values indicates a skill for predicting the quality of a generated visual artwork. This number represents the error in consistency. For example, if each of the rating pairs (Prompt vs Artwork) had the same rating, this error would be equal to 0. If each were pair was different by 1 for each rating, this error would be equal to 1, because it is an average. We asked participants to self-report their familiarity with art by asking them to self-rate their familiarity and recent visits to art shows or museums. Thus, we take the sum of both questions (which each range from 1 to 5) and test for correlation to the participants' rating scores. The test shows a weak ($r = -.31$, 95% CI [-.54, -.04]) but significant (p = .02) negative correlation between the average error in consistency and art experience. This indicates that with more reported art experience, the difference in ratings of Prompts and visual Artworks may decrease (see Figure 4).

4 STUDY 2: WRITING PROMPTS

Our aim with this study was to probe laypeople' ability to come up with effective input prompts for text-to-image systems with a specific focus on generating digital artworks. The presence of style modifiers in an input prompt may indicate an understanding of text-to-image generation and how effective prompts can be formulated.

4.1 Method

4.1.1 Study Design. We designed a creative crowdsourcing task eliciting three textual prompts from each participant. The task included a short introduction and the following instructions:

> *Imagine an artificial intelligence that turns textual input prompts into digital artworks. Your task is to produce three artworks. To this end, you will write three different input*

prompts for the artificial intelligence. You should aim to maximize the visual attractiveness and aesthetic qualities of the digital artworks generated from your input prompts.

We did not mention that prompt modifiers could be used in the prompt and wrote the instructions to avoid priming participants with a specific style (i.e., we told participants to produce 'artworks' rather than 'paintings'). Participants were asked to make their artworks as visually attractive and high-quality as possible. Note, however, that we did not aim to precisely measure attractiveness and quality, but wanted participants to think about the overall visual and aesthetic quality of the images. Participants were told that there was no right or wrong answer, but tasks would be rejected if they didn't follow the instructions.

As additional questions in this task, we asked whether the participant had experience with text-to-image generation and we collected basic demographics. Participants were paid US$0.16 per completed task. The pricing was estimated from the average task completion times in a pilot study (N = 10, US$0.12/task). In this pilot study, we noticed some participants wrote a series of consecutive instructions for the AI. The task design and instructions waswere subsequently adjusted to elicit complete prompts.

4.1.2 Participant Recruitment. We recruited 137 unique participants from Amazon Mechanical Turk using the same qualification criteria as in Study 1. Ten tasks had to be rejected due to clearly no attempt being made to answer the task with relevant information. The ten tasks were republished for other participants. After collecting the data, we manually reviewed the results and removed a further twelve responses from participants who obviously tried to game the task. The final set includes 375 prompts written by 125 unique participants (three prompts per participant).

4.1.3 Analysis. The analysis of the prompts was conducted with mixed methods. For each prompt, we qualitatively and quantitatively analyzed the prompts, as follows.

Prompt modifiers. We analyzed whether the prompts contained certain keywords and phrases commonly used in the AI art community to modify the style and quality of AI generated images [38, 39]. We decided on manual analysis because a preliminary screening revealed that very few prompts contained prompt modifiers. Each prompt was analyzed by an author of this paper. We coded the presence of prompt modifiers and report on their nature and use. We did not calculate inter-rater agreement because the coding was straight-forward [35].

Descriptive language. A prompt written in descriptive language is likely to generate images of high quality. We quantitatively assessed whether the prompts contained descriptive language by calculating a number of statistical indices for each prompt:

- The number of words (tokens) and unique words (types) in the prompt. In general, longer prompts are more likely to include certain keywords (whether on purpose or by accident) that may trigger the image generation system to generate images with high quality or in a certain style.
- The Type-Token Ratio (TTR) [23], a standard measure for lexical diversity defined as the number of types divided by the number of tokens in the prompt.[4] A token, in this case, is a discrete word whereas a type is a unique token in the prompt. For calculating the TTR, we used Kristopher Kyle's lexical-diversity Python package [29].

[4]We also experimented with other indices of lexical diversity, such as the Moving-Average Type-Token Ratio (MATTR) [7] and the Measure of Textual Lexical Diversity (MTLD) [34]. However, these measures highly depend on the text length [53]. Only a small fraction of the prompts in our sample meet the recommended minimum number of tokens for applying lexical diversity measures [57]. The use of lexical diversity indices, such as the TTR, for comparing texts of different size is not recommended [53]. In our study, we do not use the TTR for comparing the lexical diversity of prompts, but to assess the amount of repetition of tokens in the prompt.

We further used tokenization and parts-of-speech tagging (with the Natural Language Toolkit [4]) to calculate the amount of:

- Nouns (NN) and verbs (VB) as an indicator of the subjects in the prompt.
- Adjectives (JJ) as an indicator of descriptive language.
- Prepositions (IN) as an indicator of specific language.
- Cardinal numbers (CD) as information on the number of subjects.

Each prompt typically contains at least one noun as the main subject. Using descriptive and specific language is likely to improve the outcome of image generation. However, overusing prepositions could result in low fidelity of the image to the prompt. Cardinal numbers are important for prompt writing because a determinate number of subjects (e.g. *"two horses"*) is likely to provide higher-quality images than an indeterminate number (e.g. *"horses"*).

4.2 Results

4.2.1 Participants. The 125 participants in our sample included 55 men, 67 women, 1 non-binary, and two participants who did not to disclose their gender identity. The age of participants ranged from 19 to 71 years (M = 41.08 years, SD = 13.44 years). The majority of participants (98.40%) reported English being their first language. Thirty-seven participants (30.33%) responded positively to the question that they had *"experience with text-based image generation systems."* We had no explanation for this surprisingly high number at this point, but inquired more about the participants' background in our follow-up study in Section 5. Median completion times were higher than estimated in the pilot study, reaching 197 seconds. It is possible that completion times are skewed due to participants reserving tasks in bulk.

(a) NN (b) JJ (c) VB (d) IN (e) CD

Fig. 5. Wordclouds created from the 375 prompts provided by 125 workers in Study 3, split into nouns (NN), adjectives (JJ), verbs (VB), prepositions (IN), and cardinal numbers (CD). English stop words have been removed in this figure.

4.2.2 On the use of descriptive language. The prompts were of varying length, ranging from 1 to 134 tokens with an average of 12.54 tokens per prompt (SD = 14.65 tokens). Overall, the length of prompts was appropriate for text-to-image generation with only four participants producing overly long prompts. On average, participants used 3.27 nouns to describe the subjects in their prompt (SD = 3.36). Participants used verbs only sparingly in their prompts (M = 0.36, SD = 1.02). The average number of prepositions (M = 1.78, SD = 2.27) was higher than the average number of adjectives (M = 1.65, SD = 1.95). However, this number is skewed by four participants who

provided long prompts. These participants were very specific in what their images should contain, with many prepositions being used to denote the relative positions of subjects in the artwork ($Max = 21$ prepositions per prompt).

Overall, participants used rich descriptive language. The participants were creative and often described beautiful natural scenery. The main topics in the participants' prompts were landscapes, sunsets, and animals (see Figure 5). We note that the richness of the language in the prompts primarily is a result of the use of adjectives (see Figure 5b). On average, participants used 1.65 adjectives in their prompt ($SD = 1.95$). Colors, in particular, were popular among participants to describe the subjects in their artworks. The following prompts exemplify the creativity and the use of descriptive language among participants:

> *beautiful landscape with majestic mountains and a bright blue lake*
> *bright yellow sun against a blue sky with puffy clouds*
> *A fruit bowl with vibrant colored fruits in it and a contrasting background*
> *A white fluffy puppy is playing in the grass with a large blue ball that is twice his size.*
> *A shiny black horse with eyes like coal run in a lush green grassy field*
> *There should be a beautiful green forest, full of leaves, with dark brown earth beneath, and a girl in a dress sitting on the ground holding a book.*

More than half of the prompts (58.13%) did not repeat any tokens (that is, they had a TTR of 1; $M = 0.94$, $SD = 0.10$). Most of the repetitions in prompts stem from the participants' need to identify the relative positions of subjects in the image (e.g., *"[...] Touching the black line and going all the way across the top of the black line should be a dark green line. Above the dark green line should be a medium green line. [...]"*). Repetitions, as a stylistic element in prompts [39], were not being used. Only 27 prompts (7.2% of all prompts) contained cardinal numbers ($M = 0.07$, $SD = 0.29$). The cardinal numbers are depicted in Figure 5e. Two of the cardinal numbers refer to a period in time which could potentially trigger the image generation system to produce images in a certain style.

Even though we tried to mitigate it in the task design and the instructions, we noticed 18 participants (14.4%) still provided direct instructions to the AI instead of prompts describing the image content. These participants either wrote three separate instructions to the AI (e.g., *"Generate a white 250 ml tea glass [...],"* *"Draw three separate triangles [...],"* and *"Show me some digital artwork from a brand new artist."*) or they wrote three consecutive instructions as we had observed in our pilot study. The latter may not include nouns as subject terms and could thus result in images with an undetermined subject (e.g., *"sharpen image"*). Two participants thought they could chat with the AI, asking it, for instance, *"Which do you prefer: starry night sky or blue sea at dawn?,"* *"Enter your favorite geometric shape,"* and *"Can you paint me a rendition of the Monalisa?"*.

4.2.3 On the use of prompt modifiers. Even though participants were specifically instructed to create a digital artwork, we found only very few participants included style information in their prompts. Many participants described a scene in rich descriptive language, but neither mentioned artistic styles, artist names, genres, art media, nor specific artistic techniques. The participants' prompts may have described an artwork, but without style information, the style of the generated image is left to chance and the resulting images may not match the participants' intent and expectations.

Overall, the prompts did not follow the prompt template mentioned in Section 2.3 and best practices common in the AI art community were not followed. Only one participant made purposeful use of a prompt modifier commonly used in the AI art community. This prompt modifier is *"unreal*

engine."[5] The participant used this modifier in all her three prompts by concatenating it to the prompt with a plus sign, e.g.*"rainbow tyrannosaurus rex + unreal engine."* A small minority of participants used generic keywords that could trigger a specific style in text-to-image systems. For instance, the generic term *"artwork"* was used in 16 prompts (4.3%). The following list of examples reflects almost the entire set of prompts containing explicit style information among the 375 prompts written by participants (with style modifiers underlined):

Cubism portrait of a Labrador Retriever using reds and oranges
Paint a portrait of an old man in a park.
Draw a sketch of an airplane.
Abstract trippy colorful background
surreal sky castle
Can you paint me a rendition of the Monalisa?
Bob Ross, Claude Monet, Vincent Van Gogh
Are you able to produce any of rodans work.
what can you do, can you make pointillism artwork?

Besides this sparse — and sometimes accidental — addition of style information, we find that overall, participants did not control the style of their creations. Output styles were mainly determined by the participants' use of descriptive language.

5 STUDY 3: IMPROVING PROMPTS

In a follow-up study, we investigated whether participants could improve their artworks. This study aimed to answer the question of whether prompt engineering is a skill that we humans apply intuitively or whether it is a learned skill that requires expertise (e.g., by learning to write prompts from repeated interactions with the text-to-image system) and knowledge of certain keywords and key phrases (prompt modifiers), as discussed in Section 2.3 and Section 4.2.3. We hypothesize that if prompt engineering is a learned skill, participants will not be able to significantly improve their artworks after only one iteration.

5.1 Method

5.1.1 Study Design. We invited the same participants who participated in Study 2 to review images generated from their own prompts. Participants were then asked to improve their three prompts. To this end, we designed a task that introduced the participant to the study's purpose, using the same instructions as in the previous study. We additionally highlighted that if the images presented to the participant did not look like artworks, the prompt should be adjusted accordingly. Like in the previous study, we avoided to mention that prompt modifiers could be used to achieve this aim.

Participants were given five images for each of the three prompts they wrote in Study 2. We used the workerId variable on MTurk to load the participant's previous prompts and images. Participants were then asked to rewrite and improve their three prompts. The task included two input fields, one pre-filled with their previous prompt and one for optional negative terms. In practice, negative terms are used by prompt engineers to control image generation. For example, adding *"watermark"* or *"shutterstock"* to a prompt can reduce the occurrence of text and watermarks in the output. We studied this by incorporating it into our study design. Participants were introduced to the potentially surprising effects of negative terms with an example. The example explained that adding

[5]The long-form of this modifier is *"rendered in UnrealEngine,"* a computer graphics game engine. Images generated with this prompt modifier may exhibit increased quality due to photo-realistic rendering.

"zebra" as a negative term to a prompt for a pedestrian crossing could potentially result in an image of a plain road (due to stripes being removed).

For each prompt, we also collected information on whether the images matched the participant's original expectations (given the previous prompt) and whether the participant thought the prompt needed improvement (both on a Likert-scale from 1 – Strongly Disagree to 5 – Strongly Agree). The latter was added to identify cases in which participants thought that no further improvement of the prompt was necessary. We also asked participants to rate their confidence that the new prompt would result in a better artwork (on a Likert scale from 1 – Not At All Confident to 5 – Highly Confident). The task concluded with demographic questions, including the participant's experience with text-based image generation and interest in viewing and practicing art. The task design was tested and improved in a small-scale pilot study (N = 8; US$1 per task). The payment was set to US$1.75, aiming for an hourly pay of above minimum wage in the United States.

5.1.2 Research Materials. In this section, we describe how we selected an image generation system and how we generated images from the participants' prompts.

System selection. We experimented with different text-to-image generation systems, including CLIP Guided Diffusion (512x512, Secondary Model)[6], CLIP Guided Diffusion (HQ 512x512 Uncond)[7], DALLE-E mini[8], Disco Diffusion 5.3 and 5.4[9], Latent Diffusion[10], and Majesty Diffusion 1.3[11]. In the end, we selected Latent Diffusion for two main reasons. Latent Diffusion is the foundation for many of the community-driven adaptations and modifications. More importantly, the system is deterministic and leads to reproducible outcomes. Consecutive runs with the same seed value will generate the same images. This is a crucial requirement since we aim to compare images in between studies.

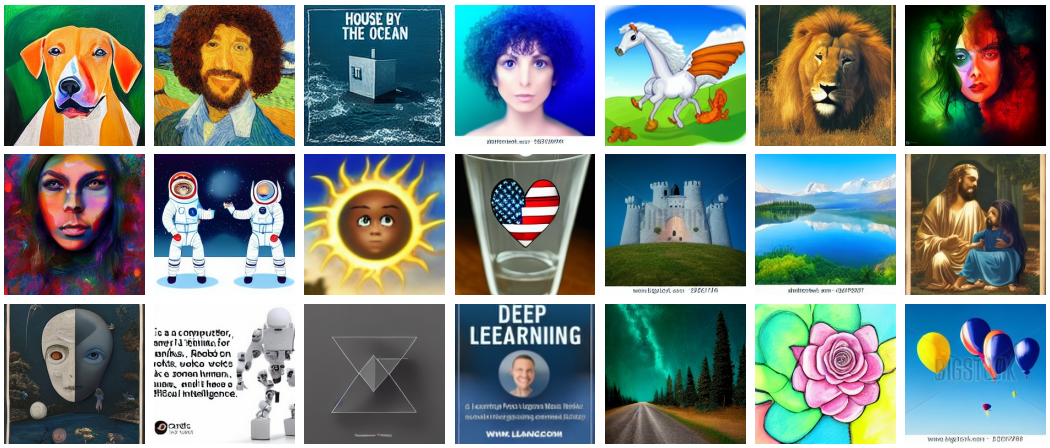
Image generation. We generated images for the participants' prompts with Latent Diffusion using the following configuration settings: text2img-large model (1.4B parameters), seed value 1040790415, eta 1.0, ddim steps 100, and scale 5.0. Even though the system is capable of generating images at higher resolutions, we decided to generate images of 256 × 256 pixels to avoid the quirks that often occur when generating images in resolutions that the model was not trained on. The image generation job yielded 1875 images (125 participants × 3 prompts per participant × 5 images per prompt). After collecting the revised prompts from participants, we generated another set of 1875 images using the same seed value and configuration settings as before. Negative terms were used in this second set, if provided by the participant.

Some hand-selected images generated from the prompts are depicted in Figure 6. Many images were of photo-realistic quality, depicting landscapes, sunsets, beaches, and animals. Besides photographs, artistic styles included paintings, graphic designs, abstract artworks, as well as magazine and book covers. Some images contained text and many images contained watermarks.

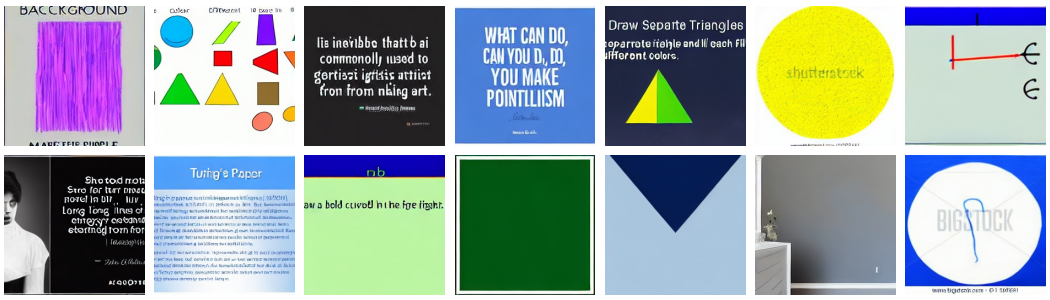
5.1.3 Analysis. We analyzed the two sets of prompts and images written in studies 2 and 3 as follows.

Analysis of prompts. To measure the amount of changes in the prompts, we calculated the number of tokens added and removed using parts-of-speech tagging as well as the Levenshtein

[6]https://colab.research.google.com/drive/1mpkrhOjoyzPeSWy2r7T8EYRaU7amYOOi
[7]https://colab.research.google.com/drive/1QBsaDAZv8np29FPbvjffbE1eytoJcsgA
[8]https://github.com/borisdayma/dalle-mini
[9]https://github.com/alembics/disco-diffusion
[10]https://github.com/CompVis/latent-diffusion
[11]https://github.com/multimodalart/majesty-diffusion



(a)



(b)

Fig. 6. Selected exemplars of a) successful and b) failed image generations from worker-provided prompts. The images in Figure 6a were selected to represent a variety of different styles and are not representative of the whole set of images. The images in Figure 6b depict some of the recurring issues in images generated from worker-provided prompts.

distance [31], a measure of lexical similarity denoting the minimum number of edits needed to change one string into another. To understand the nature of the changes, the first author inductively developed a coding scheme [20] with eight categories: adjectives/adverbs, subjects, prepositions, paraphrasing/synonyms, reordering, cardinal numbers, simplification, and presence of prompt modifiers. After discussing the codes among all authors and revising the codes, the first author coded all prompts and generated a co-occurrence matrix of changes made by participants. Note that we understand "subjects" in the sense of subject terms [39] for image generation (e.g. "a woman holding a phone" would have two subjects (woman and phone). Synonyms were analyzed at the level of individual words and parts of sentences.

Analysis of the revised images. We evaluated the images as follows. One author first created a spreadsheet with the two prompts and respective sets of five images from studies 2 and 3. The authors then discussed 30 of these image-text pairs and developed a set of evaluation criteria. Each author individually rated 50 pairs of images along these criteria. After the initial round of coding, the authors discussed the results and decided to add four more criteria to the coding scheme. The

final set of criteria included binary categories for failed generations, amount of style and subject change, and whether consistency improved, as well as ratings for details, contrast, color, distortions, watermarks, and overall subjective impression of quality. After a second round of coding, the authors cross-checked their evaluations and resolved differences through discussion.

5.2 Results

5.2.1 Participants. The sample consisted of 50 crowd workers (40% of the participants who participated in Study 2). Participants included 25 men, 24 women, and 1 person who preferred not to disclose the gender identity, aged 20 to 71 years (M = 42.76, SD = 14.63). Participants came from varied educational backgrounds, including some completed college courses (17 participants), Bachelor's degrees (22 participants), Master's degrees (4 participants), and doctorate degrees (2 participants). Seven out of ten participants had an educational background in the arts. Some partic-

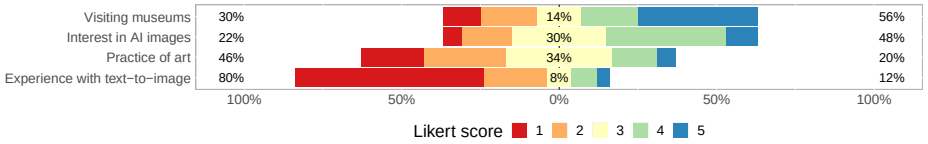


Fig. 7. Background of the crowd workers participating in Study 4.

ipants were interested in visiting museums and AI-generated imagery, but most did not practice art themselves and 80% had little or no experience with text-to-image generation.

Approximately 40% of participants were disappointed with the generated images, while 55% of participants' expectations were met. Around 60% of participants believed the images needed improvement, and a similar percentage of participants were confident that their revised prompts would improve the generated images.

5.2.2 Participants' revised prompts. The average Levenshtein distance between the participants' two prompts (not including negative terms) was 28.1 (SD = 25.0). A computational analysis of the changes with parts-of-speech tagging shows that participants added over twice as many tokens as they removed — 538 added tokens versus 243 removed tokens (see Figure 9a). Nouns were added most often (29.55% of added tokens), followed by adjectives (22.12%), prepositions (17.84%) and determiners (8.55%). The same types of tokens were also most often removed (28.81% of removed tokens were nouns, 16.87% prepositions, 13.17% adjectives, and 8.64% determiners). In 11 prompts (7.33%), the participant neither changed the prompt nor provided a negative term. Six of these instances consisted of participants pasting random snippets of text.

Table 3. Evaluation of changes in the two sets of images generated from worker-provided prompts.

	details	contrast	color	distortions	watermarks	consistency	overall
worse	17 (11.3%)	17 (11.3%)	12 (8.0%)	32 (21.3%)	31 (20.7%)	23 (15.3%)	23 (15.3%)
same	81 (54.0%)	85 (56.7%)	88 (58.7%)	99 (66.0%)	85 (56.7%)	95 (63.3%)	77 (51.3%)
better	52 (34.7%)	48 (32.0%)	50 (33.3%)	19 (12.7%)	34 (22.7%)	32 (21.3%)	50 (33.3%)

Our coding showed that the main strategy used by participants was modifying (i.e., adding, removing, or switching) adjectives in their prompts (see Figure 9b). For example, a participant changed the

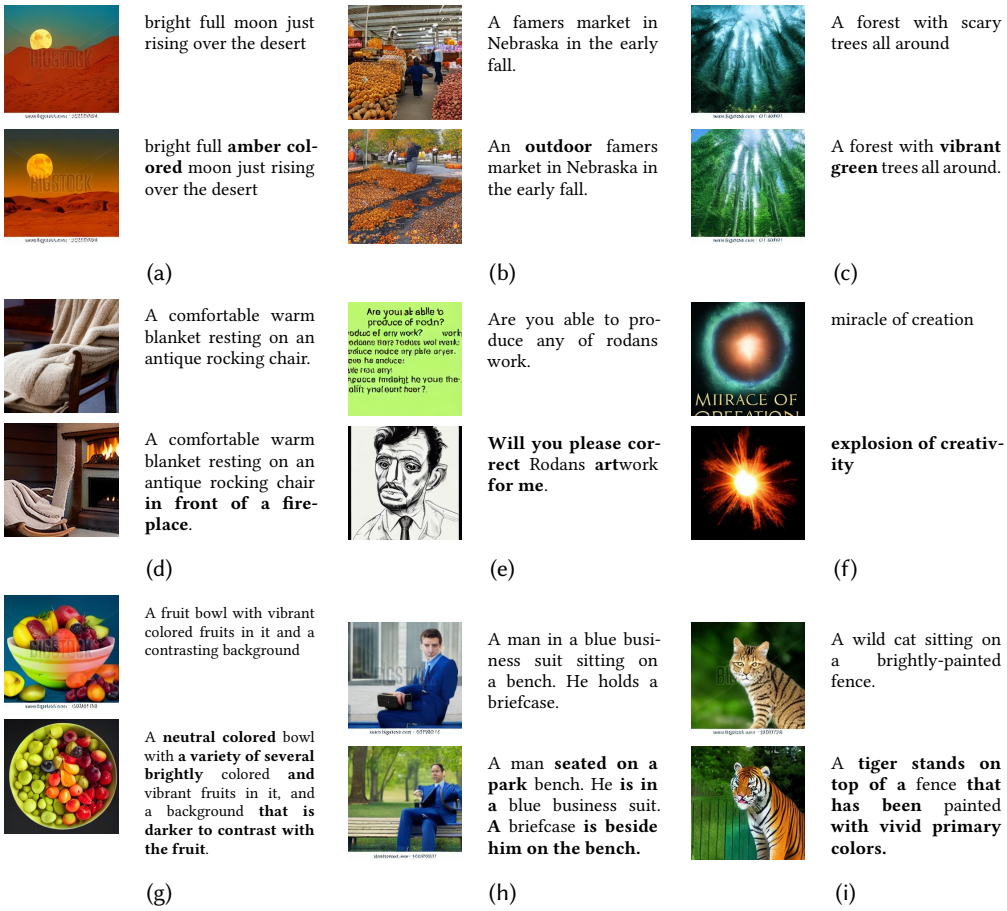
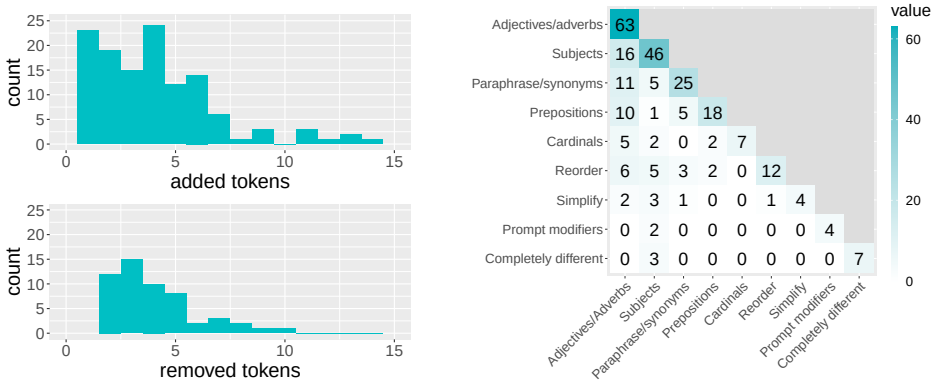


Fig. 8. Examples of changes (highlighted in bold) in adjectives and adverbs (a–c), subjects (d–f) and multiple changes at once (g–i) made by crowd workers to their own prompts in Study 4.

prompt *"flowers in winter"* to *"purple flowers in winter."* This was often combined with changes to the subject of the prompt (cf. Figure 8), such as changing *"sweeping arcs"* to *"deep and broad, sweeping arcs in landscapes."* Some participants also adapted their prompts based on what they saw in the images, though this often resulted in only minor changes to the revised images. For instance, in the case of the above participant, the two images of mountainous landscapes were almost identical. Another common approach was changing prepositions in the prompts. Few participants attempted to simplify their prompts, and relatively few made changes to cardinal numbers. For instance, one participant changed *"draw a bunch of circles"* to *"draw at least 15 circles,"* and another participant wanted to see *"lots of puffy clouds"* without specifying the exact number.

We found that only one participant (the same as in Section 4.2.3) demonstrated knowledge of prompt modifiers in all three of her prompts. An example written by this participants is *"rainbow tyrannosaurus rex, prehistoric landscape, studio ghibli, trending on artstation."* This participant used the underlined prompt modifiers which are commonly used in the AI art community. Only one other participant used a style modifier (*"real photos of [...]"*) in one prompt. This shows a very small



(a) Histogram of changes in tokens

(b) Co-occurrence matrix of changes

Fig. 9. Histograms of changes in prompt tokens and co-occurence matrix of changes made by crowd workers to their prompts in Study 4.

increase in the use of prompt modifiers among participants in between Study 2 and Study 3, even though participants were specifically instructed to improve their artworks.

5.2.3 Participants' revised images. We compared the two sets of images generated from each participant's prompts and found that over half of the revised sets showed no improvement in image quality (in terms of details, contrast, color, distortions, watermarks, and consistency). Selected changes in the prompts and the resulting images are depicted in Figure 8. About half of the sets remained the same, 15% were worse, and a third were better compared to the previous set.

Some participants were able to make improvements to the generated images, mainly by adding more details. Since participants added more tokens than they removed, the prompts were longer and resulted in about a third of the images having more details. Some participants also improved the images' colors and contrast by adding adjectives to the prompts. For instance, one participant improved the amount of details by adding *"coral reef"* to the end of the prompt *"scuba diver exploring unknown ocean."* This change resulted in less blur and more details in the coral reef. However, strong changes in the style of the images were rare, with about 70% of the revised sets being in the same or very similar style. Because participants did not use style modifiers, the revised images often resembled the initial images.

About 15% of the images were of low aesthetic quality, often consisting of text with no discernible subject (see Figure 6b). These images were rarely improved between the studies, and when they were, it was often due to chance. For instance, the subject was completely changed in about 10% of the images. This was often a result of participants trying to have a conversation with the AI and entering a completely different prompt as input (see Figure 6b and Figure 8e).

5.2.4 Participants' use of negative terms. Nineteen participants used negative terms, with eight using them in all three prompts. In total, we collected 39 negative terms. Many negative terms ($n = 19$) aimed at removing or modifying the subject in various ways, such as removing *"rocks"* from a beach, trying to correct a *"weird face,"* avoiding a *"Nude, Naked, White, Man,"* removing the color *"Green."* in the image of a red star, or attempting to change the subject entirely (*"ballroom"*). Participants tried to change the style of the images in eight cases, using terms such as *"Black, White, Colorless, Monochromatic," "opaque, solid,"* and *"unfocused."* Four out of the 50 participants tried to remove text in the images, using negative terms such as *"letters," "captions,"* and *"text."* Only one

participant attempted to remove watermarks, using the negative term *"remove watermark."* As can be seen in this prompt, some participants did not understand the concept of negative terms, even though we explained it to them. A few examples of failed and successful attempts are depicted in Figure 10. Some of the image generations failed, because the participant did not use the negative term correctly. For instance, the prompt on the bottom right of Figure 10 contains a monarch butterfly both in the prompt and negative term. The resulting image is sub-par compared to the image generated from the participant's original prompt.

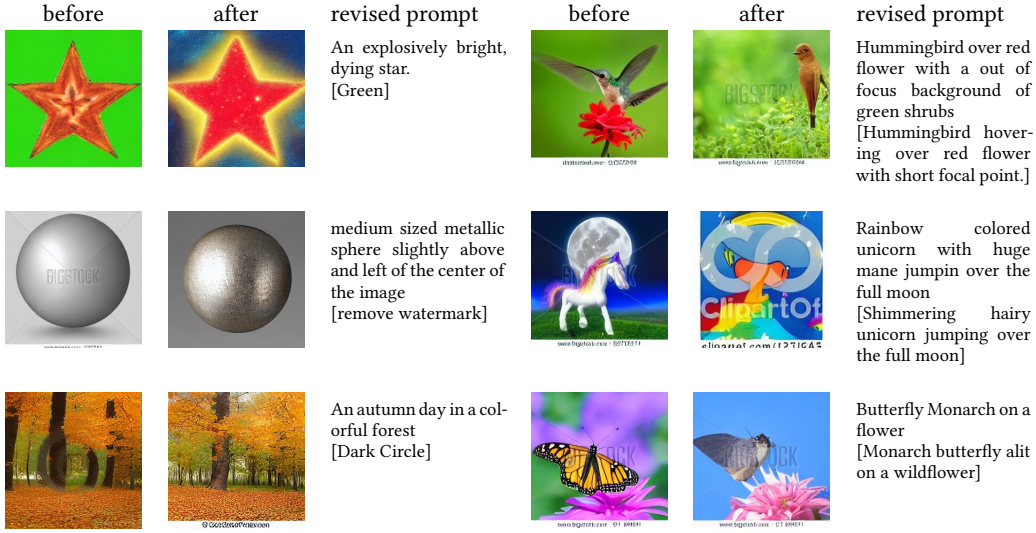


Fig. 10. Examples of successful (left) and failed (right) attempts of participants' using negative terms (in square brackets).

6 DISCUSSION

In three studies, we explored the skill of prompt engineering with participants recruited from a crowdsourcing platform. Our first study shed light on whether people have an understanding of what makes a "good" prompt. Interestingly, participants were better able to identify the quality of prompts than images. This is surprising, given the difficulty of the task of imagining the visual outcome of a prompt. The presence of style modifiers may have influenced participants to rate these prompts higher than prompts without style modifiers. Our findings nevertheless indicate that participants can assess the quality of prompts and respective images. This ability increased with the participants' experience and interest in art.

In the subsequent two studies on writing and improving prompts, we found that many participants were creative and described artworks in rich descriptive language which may result in beautiful digital artworks. However, only a negligible amount of participants intentionally used terms commonly applied in communities of text-to-image art, such as Midjourney and Stable Diffusion. With the exception of one participant, these specific keywords are not (yet) part of the vocabulary of crowd workers on MTurk. While the prompts written by participants were very descriptive and, in some cases, resulted in beautiful and interesting images, participants left the style of their image to chance. The prompts were missing modifiers that would more tightly control the style and quality of the image generation. This applies to both the initial study on writing prompts and

the study on revising and improving prompts. In the latter, only a minority of the participants were able to improve their revised images, while most of the images remained about the same quality. An overwhelming majority of participants left the style of the generated images to chance, even though they were instructed to create "artworks."

6.1 Prompt Engineering as a Learned Skill

Our work adds to the discussion of broader research questions in the AI research community: Can anyone become an artist with prompt engineering? Is prompt engineering a skill that is intuitive to us humans or is it a learned skill? How steep is the learning curve to prompt engineering? If prompt engineering is an intuitive skill, how intuitive is it? These research questions have implications for the future of work and human-computer co-creativity [25]. If prompt engineering is an intuitive human skill, then we can look forward to a bright future where anybody can work in creative professions without having to develop special skills. But if prompt engineering is a learned skill, its application could become limited to highly trained and skilled class of professionals who have mastered to speak the language of the generative model. The latter case would clearly negatively impact creative production and stifle innovation.

Prompting is a language-based practice and the use of language is intuitive to us humans. Therefore, one could assume that prompting is an intuitive skill. It is easy to get started with writing prompts and prompting has a large potential in different fields and for many application domains. However, our study found that effective prompt writing requires knowledge of keywords and key phrases. These prompt modifiers are an essential part of the skill of prompt engineering for AI generated art. Typically, these keywords and key phrases are acquired through iterative experimentation and by learning from prompts shared in dedicated resources, on social media, or in online communities [38]. Our studies empirically confirm that style modifiers are unknown to participants recruited on Amazon Mechanical Turk. Prompt modifiers that are being used profusely in the AI art community have not found their way into the collective vocabulary of workers on MTurk. Participants in our study struggled to write and improve their prompts for the specific task of creating digital artworks. This points towards prompt engineering being a learned skill or perhaps even a specialist skill, as we discuss in the following section.

6.2 On the Future of Creative Production with Prompt Engineering

Text-to-image generation opens new opportunities for creative production of digital images and artworks. Whether prompt engineering will become an expert skill or even a novel profession is still open. In this section, we speculate on four possible futures of prompt engineering.

6.2.1 Prompt engineering as an expert skill. In the future, prompt engineering could become an expert skill that requires deep subject-matter expertise (e.g. knowledge of keywords and prompt modifiers, but also of the training data and system configuration settings) to effectively control the output of generative systems. This is similar to the move in the field of machine learning towards "foundation models" [5]. Foundation models are very large and costly to train, operate, and maintain. As a result, research on these models is limited to a small number of well-financed research institutes that employ highly-skilled professionals. If prompt engineering becomes a highly skilled profession, it may become exclusive to a narrow group of privileged individuals who have undergone extensive training.

6.2.2 Prompt engineering as an everyday skill. In the future, prompt engineering could become a common practice. In this scenario, people would adapt their creative practices and language to facilitate effective interaction with AI because it is a skill that is needed in everyday life. People have a need for visual content, and AI-generated content could satisfy this need, from internet

memes to the design of greeting cards, logos, and artworks. Prompt engineering could also be used for self-actualization, creativity, and art therapy to improve mental health and well-being [6]. In this scenario, people would expand their vocabulary to include terms used in prompt engineering in order to produce meaningful outcomes with generative systems. Learning prompt engineering would be similar to learning a new language, and it could even become part of media literacy education in schools.

6.2.3 Prompt engineering as an obsolete skill. In the future, prompt engineering could become irrelevant. Prompt engineering can be seen as the smell of a half-baked product that does not solve its users' needs. As generative systems improve their ability to understand the intent of users, prompt engineering could be a short-lived trend. The problem of aligning AI with human intent is known as AI alignment in the scholarly literature [13]. State-of-the-art systems, such as ChatGPT [37] and DALL-E 2 [46], demonstrate impressive performance in understanding textual input prompts and user intent. With these systems, users of all skill levels can generate content from textual prompts. As generative systems become better at understanding input prompts, prompt engineering could become unnecessary — an archaic skill that does not require expert training and that only few people exercise for nostalgic reasons. In this scenario, the machine would adapt to humans, rather than the other way around. The generative machine would develop an "intimate" relationship [55] with its users and a perfect understanding of user intent.

6.2.4 Prompt engineering as personal signature or curation skill. In the future, our human senses could become better at distinguishing hand-crafted art from AI-generated digital art. AI artist Mario Klingemann speculated that with the influx of AI-generated images, this skill would help us notice subtle nuances, details, and imperfections in AI-generated art, which could become more important in determining the aesthetic quality of an art piece.[12] In this scenario, anyone could write prompts for generative systems with good results, but only a few would become masters of prompt engineering. The practice of prompt engineering would remain a necessary skill for applying finishing touches and optimizing generative results, as well as imbuing an artwork with a personal style to distinguish it from bland "off-the-shelf" generations. Alternatively, prompt engineering could evolve into a curation skill — a personal practice in which everyone has their own curated sets of textual and visual input prompts used to fine-tune generative models for different purposes. Current systems that cater towards this future are Dreambooth [49] and the method of textual inversion [14].

In the following, we formulate recommendations for involving paid crowd workers in research on text-to-image systems.

6.3 Recommendations for Conducting Text-to-Image Experiments with Crowd Workers

In this section, we reflect on our experiences with conducting experiments on prompt engineering with a paid crowd on Amazon Mechanical Turk. Our research has found that crowd workers are capable of coming up with creative prompts written in descriptive language. However, it is important for workers to receive training in prompt engineering in order to gain a better understanding of how to use prompt modifiers. We have put together ten recommendations for conducting experiments on text-to-imagine generation with a paid online crowd.

6.3.1 Recruit the right crowd. Some workers on crowdsourcing platforms enjoy and even seek out creative crowdsourcing tasks [40]. Workers who intrinsically enjoy writing stories or other types of creative crowdsourcing tasks are likely to produce better results with text-to-image generation

[12] https://twitter.com/quasimondo/status/1512769106717593610

systems. We see that in some of the workers in our sample who invested effort into writing detailed and descriptive prompts, while others consistently produced images of low aesthetic quality. It is therefore important to recruit the right crowd, for instance by using qualification tasks.

6.3.2 Provide guidance and clear task instructions. Workers in our studies left the style of the generated images to chance, even though they were instructed to create "artworks." Therefore, the task instructions should include clear and detailed explanations of what is expected from the worker, and how the worker can achieve these goals. This includes explaining what prompts are, how they should be phrased, and how to correctly include prompt modifiers in prompts.

6.3.3 Explain keywords and key phrases to workers. Only a few workers in our studies intentionally used prompt modifiers. Keywords and key phrases used in the AI art community were virtually unknown to workers (with one exception). While we expect this situation to improve as the popularity of text-to-image systems increases, workers will need to receive explanations and examples of prompt modifiers for workers to be able to use them. Task instructions should be illustrated with exemplars of prompts and resulting images to educate workers before they write their first prompt.

6.3.4 Be mindful of lengthy task instructions. While detailed instructions are needed, long instructions may slow down workers. Some workers may even ignore long instructions. An alternative to lengthy instructions is to design instructions as a playful tutorial to train workers. Another approach would be to use the idle time between image generations for providing useful information to workers. OpenAI's DALL-E interface, for instance, provides helpful tips while one waits for the image generation to finish. This principle could be applied to crowdsourcing tasks as well.

6.3.5 Explain negative prompts. Some workers in our sample did not understand the concept of negative terms, even though we explained it to them and provided an illustrative example. Crowdsourcing tasks should include a carefully designed explanation of how negative terms work, ideally with example images that demonstrate the effect of negative terms.

6.3.6 Create dedicated user interfaces. Instead of training workers in the use of prompt modifiers, it may be more effective to provide them with user interfaces that hide the complexity of open-domain prompting. One example of this is MindsEye[13], which is built on top of Colab notebooks and is designed with usability in mind. With this interface, even novice users can easily change image styles by simply clicking buttons, without needing to understand the underlying mechanics of prompt modifiers. This can help make the process of generating images from text more accessible and user-friendly.

6.3.7 Provide means for fast and iterative image generation. In order to effectively facilitate the process of text-to-image generation, it is important to carefully design the workflow for the workers. One way to do this is by using techniques like remixing and forking, which can help simplify the process of prompt engineering for the workers. For instance, the AI art tool Artbreeder[14] makes it easy for users to adjust image generation settings and to create variations and fork other users' images. This can help streamline the process and make it more efficient for workers.

6.3.8 Provide visual feedback to workers. Workers in our studies wrote the first set of prompts without feedback. Immediate feedback would likely improve the workers' performance in subsequent image generations.

[13] https://multimodal.art/mindseye
[14] https://www.artbreeder.com

6.3.9 Allow a training period in which workers can experiment with the text-to-image system. When workers are assigned batches of microtasks, they may need some time to understand what they need to do. This can lead to a "warming up" period and a training effect, where the worker's initial attempts at generating images are not as good as their later attempts. To account for this, the design of the crowdsourcing campaign should include measures to exclude the workers' first attempts from further analysis. This can help ensure that the results of the experiment are not biased by the workers' learning process.

6.4 Limitations

We acknowledge a number of risks to the validity of our exploratory studies.

Aesthetic quality assessment is a highly subjective task [24]. Many factors can affect ratings of aesthetic quality, such as personal values, personal background, the interestingness and content of the image, contrast, proportion, number of elements, novelty, and appropriateness [24, 26, 27]. We acknowledge that the task given to workers was hard, especially when it comes to imagining the visual outcome of a written prompt. We did, however, find that workers were able to tell the difference between low quality and high quality prompts.

We further acknowledge limitations in our choice of text-to-image system. Our main motivation for selecting Latent Diffusion was to select a deterministic system that produces reproducible results. This allowed us to assess the effect of changes in the prompt on the images. Note, however, that while Latent Diffusion is a powerful image generation system, it may respond differently to style keywords than CLIP-guided systems. However, only one participant used specific keywords (prompt modifiers) in our study. The second round of images was also never shown to participants. Therefore, we can safely assert that the choice of system had no effect on how participants wrote and revised their prompts.

7 CONCLUSION

The past few years have seen the rise of generative models. It is too early to tell whether this development will give birth to new professions, such as "prompt engineer." However, generative AI will deeply affect and reconfigure the fabric of our society. This opens exciting opportunities for research in HCI.

This article investigated prompt engineering for AI art. In three studies, we investigated whether novice participants could recognize the quality of prompts and their resulting images and whether participants could write and improve prompts. We found participants recruited from Amazon Mechanical Turk are creative and able to write prompts for text-to-image systems in rich descriptive language, but lacked the special vocabulary found in AI art communities. The use of prompt modifiers was not intuitive to participants, pointing towards prompt engineering being a learned skill. We provided recommendations for conducting scientific experiments on prompt engineering and text-to-image generation with participants recruited from crowdsourcing platforms. We speculated on four possible futures for prompt engineering. We hope that whatever the landscape of creative production will turn out to be in the future, it will be an inclusive creative economy in which everyone can participate in meaningful ways.

DATA AVAILABILITY STATEMENT

Data related to the study is available at
https://osf.io/bjwf4/?view_only=caf73282354643e9bfb34b3b05ef4b62

REFERENCES

[1] Chris Allen. 2022. Zippy's Disco Diffusion Cheatsheet v0.3. https://docs.google.com/document/d/1l8s7uS2dGqjztYSjPpzlmXLjl5PM3IGkRWI3IiCuK7g/edit?usp=sharing

[2] Shihoko Arai and Hideaki Kawabata. 2016. Appreciation Contexts Modulate Aesthetic Evaluation and Perceived Duration of Pictures. *Art & Perception* 4, 3 (2016), 225 – 239. https://doi.org/10.1163/22134913-00002052

[3] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. https://doi.org/10.48550/ARXIV.2202.01279

[4] Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Barcelona, Spain, 214–217. https://aclanthology.org/P04-3031

[5] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv* (2021). https://crfm.stanford.edu/assets/report.pdf

[6] Winslow Burleson. 2005. Developing creativity, motivation, and self-actualization with learning systems. *International Journal of Human-Computer Studies* 63, 4 (2005), 436–451. https://doi.org/10.1016/j.ijhcs.2005.04.007 Computer support for creativity.

[7] Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17, 2 (2010), 94–100. https://doi.org/10.1080/09296171003643098

[8] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. https://doi.org/10.48550/ARXIV.2209.01390

[9] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2022. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. https://doi.org/10.48550/ARXIV.2212.07476

[10] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing Discrete Text Prompts With Reinforcement Learning. https://doi.org/10.48550/ARXIV.2205.12548

[11] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. https://doi.org/10.48550/ARXIV.2105.05233

[12] Harmeet Gabha. 2022. Disco Diffusion 70+ Artist Studies. https://weirdwonderfulai.art/resources/disco-diffusion-70-plus-artist-studies/

[13] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. https://doi.org/10.48550/ARXIV.2208.01618

[15] Gernot Gerger, Helmut Leder, and Alexandra Kremer. 2014. Context effects on emotional and aesthetic evaluations of artworks and IAPS pictures. *Acta Psychologica* 151 (2014), 174–183. https://doi.org/10.1016/j.actpsy.2014.06.008

[16] Pentti Haddington, Noora Hirvonen, Simo Hosio, Marianne Kinnula, Jonna Malmberg, Siamak Seyfi, Jaakko Simonen, Sara Ahola, Marta Cortés Orduna, Heidi Enwald, Lotta Haukipuro, Mervi Heikkinen, Jan Hermes, Sanna Huikari, Netta Iivari, Sanna Järvelä, Outi Kanste, Lydia Kokkola, Sari Kunnari, and Kateryna Zabolotna. 2021. GenZ White Paper: Strengthening Human Competences in the Emerging Digital Era.

[17] Michaela Hoare, Steve Benford, Rachel Jones, and Natasa Milic-Frayling. 2014. Coming in from the Margins: Amateur Musicians in the Online Age. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, 1295–1304. https://doi.org/10.1145/2556288.2557298

[18] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B. Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling Creative Inspiration with Fine-Grained Functional Aspects of Ideas. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, Article 12, 15 pages. https://doi.org/10.1145/3491102.3517434

[19] Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. 2022. MetaPrompting: Learning to Learn Better Prompts. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3251–3262. https://aclanthology.org/2022.coling-1.287

[20] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[21] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-Based Prototyping with Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. https://doi.org/10.1145/3491101.3503564

[22] "Johannezz". 2022. The Promptist Manifesto. https://deeplearn.art/the-promptist-manifesto/

[23] Wendell Johnson. 1944. Studies in language behavior 1: A program of research. *Psychological Monographs* 56 (1944), 1–15.

[24] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z. Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and Emotions in Images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115. https://doi.org/10.1109/MSP.2011.941851

[25] Anna Kantosalo, Prashanth Thattai Ravikumar, Kazjon Grace, and Tapio Takala. 2020. Modalities, Styles and Strategies: An Interaction Framework for Human–Computer Co-Creativity. In *Proceedings of the Eleventh International Conference on Computational Creativity*, Amílcar Cardoso, Penousal Machado, Tony Veale, and João Miguel Cunha (Eds.). Association for Computational Creativity, Portugal, 57–64.

[26] Shahabeddin Khalighy, G. Green, Christoph Scheepers, and Craig Whittet. 2014. Measuring Aesthetic in Design. In *Proceedings of the DESIGN 2014 13th International Design Conference*. The Design Society, 2083–2094.

[27] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 662–679.

[28] Logan Kugler. 2021. Non-Fungible Tokens and the Future of Art. *Commun. ACM* 64, 9 (aug 2021), 19–20. https://doi.org/10.1145/3474355

[29] Kristopher Kyle. 2018. lexical-diversity Python package. https://github.com/kristopherkyle/lexical_diversity

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3045–3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

[31] Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady* 10 (1965), 707–710.

[32] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. arXiv:2101.00190 [cs.CL]

[33] Vivian Liu and Lydia B. Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, Article 384, 23 pages. https://doi.org/10.1145/3491102.3501825

[34] Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42, 2 (2010), 381–392. https://doi.org/10.3758/brm.42.2.381

[35] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. https://doi.org/10.1145/3359174

[36] Midjourney. 2022. Midjourney.com. https://www.midjourney.com

[37] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/

[38] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *25th International Academic Mindtrek Conference (Academic Mindtrek 2022)*. Association for Computing Machinery, New York, NY, USA, 192–202. https:

//doi.org/10.1145/3569219.3569352
[39] Jonas Oppenlaender. 2022. A Taxonomy of Prompt Modifiers for Text-To-Image Generation. https://doi.org/10.48550/ARXIV.2204.13988
[40] Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on Paid Crowdsourcing Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, Article 548, 14 pages. https://doi.org/10.1145/3313831.3376677
[41] Jonas Oppenlaender, Aku Visuri, Ville Paananen, Rhema Linder, and Johanna Silvennoinen. 2023. Text-to-Image Generation: Perceptions and Realities. In *Workshop on Generative AI in HCI (CHI '23')*. 5 pages.
[42] Matthew Pelowski, Gernot Gerger, Yasmine Chetouani, Patrick S. Markey, and Helmut Leder. 2017. But Is It really Art? The Classification of Images as "Art"/"Not Art" and Correlation with Appraisal and Viewer Interpersonal Differences. *Frontiers in Psychology* 8, Article 1729 (2017), 21 pages. https://doi.org/10.3389/fpsyg.2017.01729
[43] Margaret H. Pinson, Lucjan Janowski, Romuald Pepion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. 2012. The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study. *IEEE Journal of Selected Topics in Signal Processing* 6, 6 (2012), 640–651. https://doi.org/10.1109/JSTSP.2012.2215306
[44] Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art. In *Creativity and Cognition (C&C '22)*. Association for Computing Machinery, New York, NY, 15–28. https://doi.org/10.1145/3527927.3532792
[45] 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. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763.
[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/ARXIV.2204.06125
[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. https://doi.org/10.48550/ARXIV.2102.12092
[48] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. https://doi.org/10.48550/ARXIV.2102.07350
[49] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. https://doi.org/10.48550/ARXIV.2208.12242
[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. https://doi.org/10.48550/ARXIV.2205.11487
[51] Ernestasia Siahaan, Judith A. Redi, and Alan Hanjalic. 2014. Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. 245–250. https://doi.org/10.1109/QoMEX.2014.6982326
[52] Ethan Smith. 2022. A Traveler's Guide to the Latent Space. https://sweet-hall-e72.notion.site/A-Traveler-s-Guide-to-the-Latent-Space-85efba7e5e6a40e5bd3cae980f30235f
[53] Fiona J. Tweedie and R. Harald Baayen. 1998. How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32, 5 (1998), 323–352.
[54] Noah N.N. Van Dongen, Jan W. Van Strien, and Katinka Dijkstra. 2016. Implicit emotion regulation in the context of viewing artworks: ERP evidence in response to pleasant and unpleasant pictures. *Brain and Cognition* 107 (2016), 48–54. https://doi.org/10.1016/j.bandc.2016.06.003
[55] Mark Weiser. 1993. Some Computer Science Issues in Ubiquitous Computing. *Commun. ACM* 36, 7 (jul 1993), 75–84. https://doi.org/10.1145/159544.159617
[56] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. 2023. A Prompt Log Analysis of Text-to-Image Generation Systems. In *Proceedings of the ACM Web Conference (WWW '23)*.
[57] Fred Zenker and Kristopher Kyle. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing* 47 (2021), 15 pages. https://doi.org/10.1016/j.asw.2020.100505
[58] Joanna Zylinska. 2020. *AI Art: Machine Visions and Warped Dreams*. Open Humanities Press, London, UK.

A SET OF IMAGES USED IN STUDY 1

A.1 Images with High Aesthetic Appeal

H1: the foundations of origin, matte painting, genesis, trending on artstation, high resolution

H4: eclectic interior of the mind

H5: , ., ., matte painting, 8k cgsociety

H6: The Dude by Glenn Fabry

H2: vikings. by Dan Mumford, matte painting, Studio Ghibli

H7: fantastic wardrobe of the inner sanctuary comes to life in giant birtation of the soul

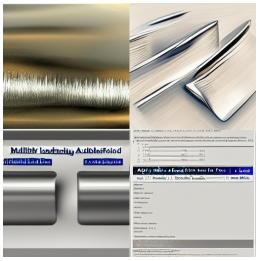
H9: tidal wave, matte painting, rendered in octane, ghibli, 8k #epic #wow trending on wikiart

H8: a moment of silence for our fallen heroes. War memorial. central. CGSociety, painting, postprocessing

H10: portrait of a world war soldier on artstation

H3: buck, Hudson River School

A.2 Images with Low Aesthetic Appeal



L1: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL

L2: a tweet about bias

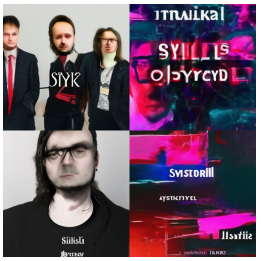
L3: Asterix at the Robot Games. by Rene Goscinny and Albert Uderzo

L4: amazing green screen effect

L5: Office Space, Bill Lumbergh. “yeah, we need you to come in on Saturday, mkay?”

L6: Blind No. 20, Seventeen-foot high Ceiling or Lower, Historical Veridian Green, Indian Yellow Hue, Hansa Yellow Medium (to Mike Kelley)

L7: we can do it! propaganda poster



L8: My New Band Is Called Syskill

L9: China buys Russia

L10: artwork, academic paper