A Taxonomy of Prompt Modifiers for Text-To-Image Generation

JONAS OPPENLAENDER, University of Jyväskylä, Finland

Text-to-image generation has seen an explosion of interest since 2021. Today, beautiful and intriguing digital images and artworks can be synthesized from textual inputs ("prompts") with deep generative models. Online communities around text-to-image generation and AI generated art have quickly emerged. This paper identifies six types of prompt modifiers used by practitioners in the online community based on a 3-month ethnographic study. The novel taxonomy of prompt modifiers provides researchers a conceptual starting point for investigating the practice of text-to-image generation, but may also help practitioners of AI generated art improve their images. We further outline how prompt modifiers are applied in the practice of "prompt engineering." We discuss research opportunities of this novel creative practice in the field of Human-Computer Interaction (HCI). The paper concludes with a discussion of broader implications of prompt engineering from the perspective of Human-AI Interaction (HAI) in future applications beyond the use case of text-to-image generation and AI generated art.

CCS Concepts: • **Human-centered computing** → *Natural language interfaces*; *Human computer interaction (HCI)*; • **Applied computing** → *Arts and humanities.*

Additional Key Words and Phrases: prompt engineering, text-to-image generation, human-AI interaction, AI generated art

1 INTRODUCTION

The past year has seen a flurry of text-to-image generation systems being developed. Based on deep learning, these systems can generate digital images from short descriptive texts (called "prompts", such as "`an old oil painting of a beautiful landscape at dawn`"). Examples of images synthesized from textual prompts are depicted in Figure 1. Given the quality of these images, it is not surprising that an online community around this novel text-based way of creating images has developed. Within this community, the practice and skill of writing prompts is known by the term "prompt engineering" due to its iterative and experimental nature [26]. Prompt engineering is an emerging research area in the field of Human-Computer Interaction (HCI) on how to phrase effective input prompts for deep generative models.

To be effective, the textual input prompts need to be given in a certain format in order to, for instance, generate images with a certain style. This is commonly achieved by adding keywords and key phrases to the prompts (so-called "prompt modifiers"). There is an emerging online ecosystem of resources and guides that help and teach novices how to write input prompts for text-to-image generation systems (e.g., [12, 15, 36, 49]). Further, a growing number of resources in the gray and scholarly literature present systematic experimentation on the effect of different prompt modifiers [12, 15, 26, 36]. However, some artists do not share their complete prompts for their AI generated artworks and it is often not clear how these artworks were created. Therefore, the learning curve can be steep and prompt engineering is a skill that is learned from extensive experimentation and trial and error [26].

Author's address: Jonas Oppenlaender, jonas.x1.oppenlander@jyu.fi, University of Jyväskylä, Jyväskylä, Finland.

XXXX-XXXX/2022/8-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

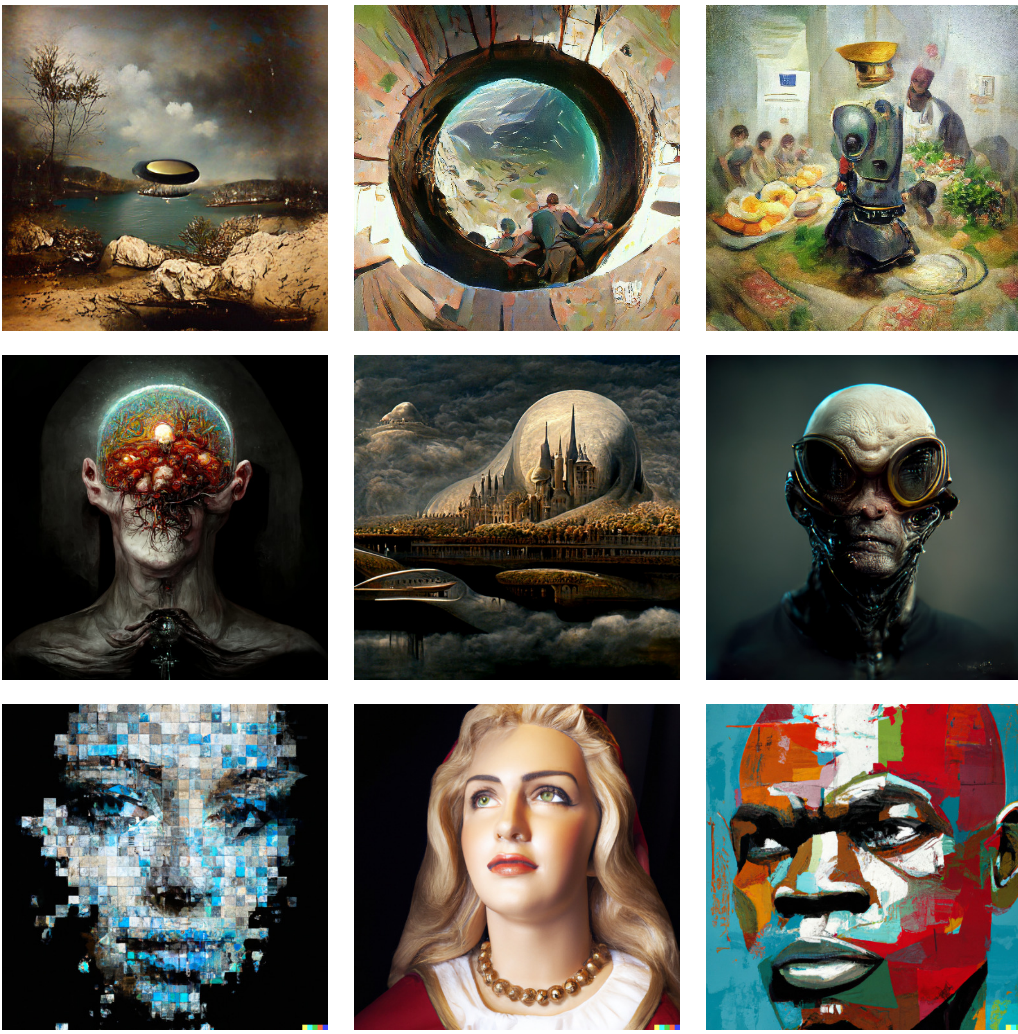

Fig. 1. Select images generated with text-to-image generation using VQGAN-CLIP (top) [8], Midjourney (middle) [27], and DALL-E 2 (bottom) [43].

No previous study has investigated different types of prompt modifiers. With a specific focus on digital art generated with text-to-image systems, this paper contributes a taxonomy of prompt modifiers used by practitioners in the text-to-image community, based on an ethnographic study of the community's prompt engineering practices on Twitter. The work is based on a three-month online ethnography which analyzed how prompt modifiers are being applied in prompt writing. This paper contributes toward a better understanding of prompt engineering as a practice within HCI in order to inform the HCI research community on the emerging practice of prompt engineering within the broader context of human interactions with artificial intelligence. This paper can enhance the theoretical understanding of how people write prompts and use prompts modifiers. Through understanding prompt writing, we can pave the way towards a broader and unified theory of prompt engineering which the HCI literature is currently missing. The paper also touches on how the technology behind text-to-image systems and the practice of prompt engineering has broader implications in research on HCI and Human-centered AI.

The paper is structured as follows. We first provide a brief introduction into text-to-image synthesis and prompt engineering in Section 2. After describing the methodological approach in Section 3, we present a taxonomy of six different types of prompt modifiers used by practitioners in the text-to-image art community (Section 5). The paper concludes with a discussion of opportunities for future research on text-to-image generation and the broader implications beyond AI generated art (Section 6).

2 BACKGROUND

2.1 Text-to-Image Generation

The field of image synthesis using deep learning has recently seen an unprecedented growth with the break-through development of multimodal models such as CLIP [41]. There has been an increase in activity in the research community in the past year and a dedicated online community around text-to-image generation with specific focus on AI generated art (which people online sometimes refer to as "AI art" [30]) has emerged. Much of the development was driven by this online community openly sharing the source code of text-to-image systems.

The development was initially spurred by OpenAI's multimodal model CLIP [41]. CLIP is a contrastive language-vision model trained in an unsupervised way that can perform zero-shot classification of images. When used as a discriminator component in text-conditioned generative systems, CLIP can "guide" the image generation process. CLIP was originally a part of OpenAI's DALL-E architecture. However, because OpenAI did not make the first version of their DALL-E system accessible to the public, the online community took over and contributed greatly to the technical development of text-to-image generation. This resulted in a vast number of text-to-image systems openly available for experimentation, first as CLIP-guided generative adversarial networks (e.g., VQGAN–CLIP [8]) and later as diffusion based image generation systems, such as CLIP Guided Diffusion [6] and Latent Diffusion [45]. In recent months, big organizations have also beta-tested text-to-image systems, such as OpenAI's DALL-E 2 [43] and Google's Imagen [47].

This paper investigates text-to-image generation from the lens of Human-Computer Interaction (HCI). In order to generate images from text, one not only has to choose the right words to make the text-to-image system generate the desired images, one also has to add different keywords and keyphrases to control the style and quality of the image generation. This creative practice of writing effective prompts is sometimes referred to as "prompt engineering." In this paper, we investigate what (and how) different types of prompt modifiers are being applied in this emerging practice.

2.2 Prompt Engineering and Prompt Modifiers

Prompt engineering [26] (or prompt design [33], prompt programming [44], or prompting for short) is an emerging practice in which "carefully selected and composed sentences are used to achieve a certain visual style in the synthesized image" [46]. The practice has seen an ideal application ground in AI generated art, but it is not limited to text-to-image generation. The term prompt engineering was originally coined to denote the practice of writing textual inputs for the autoregressive language model GPT-3 [26]. This language model requires context to produce relevant text as output. Templates have been developed to optimally provide textual inputs to GPT-3. OpenAI's documentation, for instance, lists 49 "recipes" on how to phrase input prompts for their language model [34]. Using such recipes, the output of the language model can be adapted to different down-stream tasks, such as correcting grammar, summarizing text, answering questions, generating product names, or acting as a chat bot.

Similar templates have emerged for writing input prompts for text-to-image systems, particularly in the online community around AI generated art. For instance, the "Traveler's Guide to the Latent Space" recommends the following prompt template [49]:

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

Similar templates are being followed in many resources originating from within the online community, such as the DALL-E Prompt Book [36]. Figure 2 provides an example of a typical textual input prompt and the resulting AI generated image.

Fig. 2. Digital artwork generated with DISCO Diffusion [10] from the input prompt "`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.`" This prompt is part of the default configuration settings in the DISCO Diffusion notebook.

Prompt engineering is not a hard science as found in the fields of science, technology, engineering, and mathematics (STEM). Rather, it is a term that originates from within the online community of practitioners of text-to-image generation. The term reflects the community's self-understanding, similar to the terms "AI art" and "AI artist" which also originate from within the community. Due to the rise in popularity of text-to-image systems, practitioners of AI art include not only technology-savvy developers and early-adopting hobbyists, novices, but also artists, professionals, semi-professionals, and "Pro-Ams" [22] with or without commercial interests. In the remainder of this paper, we will refer to the members of the online text-to-image community as *practitioners*.

Prompt engineering resembles a conversation with the text-to-image system. A practitioner typically will run a prompt, observe the outcome, and adapt the prompt to improve the outcome. Prompt engineering, thus, is iterative and practitioners formulate prompts as probes into the generative models' latent space. The online community quickly found that the aesthetic qualities and subjective attractiveness of images can be improved by adding certain keywords and key phrases to the textual input prompts. The terms may be referred to by a number of different names, such as "prompt modifiers," "style phrases," or "vitamin phrases" [39]. In this paper, we refer to them as *prompt modifiers*. By adding a prompt modifier to a textual input, one seeks to direct the text-to-image system in certain directions, hence "modifying" the resulting image.

In practice, prompt modifiers are applied through experimentation or based on best practices learned from experience or online resources. An example of an iterative application of prompt modifiers can be seen in Figure 3. Knowing what prompt modifiers work best for a given subject term is often the result of the practitioner's iterative experimentation, research in online communities, and the use of online tools and resources created for supporting the practice of prompt engineering [35].

a) b) c) d)

Images generated with VQGAN-CLIP [8], 175 iterations, CLIP model ViT-B/32, VQGAN model wikiart_16384, seed 6087304447281500163, text prompts:

a) "`ufo landing`"

b) "`ufo landing, daguerreotype`"

c) "`ufo landing, daguerreotype, trending on /r/art`"

d) "`ufo landing, daguerreotype, by greg rutkowski, trending on /r/art`"

Fig. 3. Example of iterative prompt engineering for generating an image.

3 METHOD

The research followed a two-fold approach. We combined autoethnographic experimentation (Section 3.1) and ethnographic research (Section 3.2), as described in this section.

3.1 Autoethnographic Research on Prompt Engineering for Text-to-Image Art

Prompt engineering is an acquired skill that is associated with a learning curve. Prompt engineering is learned through iterative experimentation akin to "brute-force trial and error" [26]. The skill can be learned from community-provided resources, such as guides, reports of systematic experimentation, or from prompts shared on social media (such as online communities dedicated to text-to-image art or Twitter). However, to appreciate and understand the craft, one has to apply the knowledge and experiment with different input prompts. In the case of text-to-image generation, autoethnography research is, therefore, an appropriate method to learn about prompt engineering.

The author conducted a 3-month autoethnographic study [9, 11, 13] between October 2021 and December 2021. This personal ethnography [5] allowed the author to get a "practitioner's perspective" [11] on text-to-image generation by "learning from self-use" [29]. The author experimented with text-to-image synthesis and created digital images with a text-to-image system using notebooks hosted on Google's Colaboratory (Colab)[1]. The author started on average at least one Colab session every work day between October 4 and December 31, 2021. The free tier of Google Colab was used in all sessions. This limited the overall working time to about 2 hours per day,

[1]https://colab.research.google.com

depending on the computational power of the assigned resources and whether penalties were incurred the previous day. VQGAN-CLIP [8] was chosen as text-to-image system using a notebook titled "VQGAN and CLIP (z + quantize method with augmentations)" [7]. This VQGAN-CLIP notebook was originally created by Katherine Crowson, with "modifications by Eleiber # 8347" and a "friendly interface" by "Abulafia # 3734" and further modifications by Justin John. VQGAN-CLIP was selected for several reasons. First, VQGAN-CLIP was one of the first text-to-image systems that experienced widespread popularity in the emerging text-to-image art community in 2021. This made VQGAN-CLIP instrumental to the growth of the community [8]. Second, the system can be executed on Google's Colaboratory (Colab) free of charge. The system requires less memory than later systems, and it is therefore less likely that image generation will fail due to insufficient memory. Third, the VQGAN-CLIP notebook on Colab is very accessible and straight-forward to use, with only a small number of configuration parameters (c.f. Figure 3). Last, the system is deterministic. Consecutive runs with the same configuration parameters will produce exactly the same images which makes the images reproducible. This is not the case with some of the later systems which make use of non-deterministic algorithms. The author generated 885 images in the course of this study.

The autoethnographic research was not conducted from scratch. Rather, it was informed by learning from the community on social media. To this end, the autoethnographic research was complemented with an online ethnography of the text-to-image art community on Twitter and a study of online community resources, described in the following section.

3.2 Ethnographic Study of the Text-to-Image Art Community

The ethnographic research of prompt engineering was conducted on the emerging text-based generative art community on Twitter and also included a literature review. The aim of the social media ethnography [37, 38] was to learn more about the textual prompts used in the community around text-to-image art. Insights derived from the study of this community were used in the autoethnographic experimentation with the text-to-image system.

3.2.1 Twitter community. During the 3-month period of research, the author took the role of "participant-as-observer" [20] by engaging with the text-to-image art community, participating in discussions, and posting digital images created with the text-to-image system on Twitter. The author followed posts on Twitter to learn about different prompts used in the text-to-image art community. To this end, the author followed trending hashtags, such as #vqganclip, #VQGAN, #clipguideddiffusion, #digitalart, #AIArt, #generativeart, and #GANArt. Not every practitioner of text-to-image art shares their prompts on Twitter. Especially if commercial interests are involved – e.g., selling the art as non-fungible token (NFTs) – practitioners may keep their prompts a secret. The research material, therefore, was sparse. However, some practitioners are more liberal in sharing their prompts. It is the posts from this group of Twitter users that informed this research (e.g., posts by Katherine Crowson (@RiversHaveWings), Hannah Johnston (@hannahjdotca), @nshepperd1, and John David Pressman (@jd_pressman), to name but a few).

3.2.2 Review of community resources. In parallel to the research on the online community, a review of the literature on the practice of prompt engineering was conducted, with specific focus on text-to-image generation of digital art. Text-to-image synthesis for AI generated art and its associated online community are a very recent phenomenon. The online community around text-to-image art only emerged in early 2021 and has seen a steady increase in activity. However, with the exception of Liu and Chilton's design guidelines for prompt engineering [26, 40], there is little scholarly literature on the practice of text-to-image generation for AI generated art in the field of HCI. Therefore, the literature review focused on sources in the gray literature, such as community-provided guides and resources, posts on blogs, and articles on the Web.

3.3 Inductive Development of the Taxonomy

The taxonomy was developed inductively from pieces of information found during the research. Due to the relative scarcity of this material, the development of the taxonomy was conducted iteratively, as follows. A list of potential candidates for prompt modifiers was inductively compiled and grouped. This list was subject to continual reinterpretation based on novel instances of prompts. Whenever a candidate for a novel type of prompt modifier was encountered in a post on Twitter or the literature, the author revisited the list of prompt modifiers. Therefore, the resulting taxonomy was iteratively and inductively revised and expanded when new types of prompt modifiers were encountered. After some weeks of collecting data this way, the list of prompt modifiers and taxonomy did no longer grow, even if instances of novel and atypical prompts were encountered. This indicates the completeness of the developed taxonomy.

The findings were documented in a PowerPoint presentation with text and images to produce an evocative and aesthetic description of the ethnographic research. This iteration also served as verification of the correctness of the taxonomy. The author's creation of and engagement with the presentation acted as a daily conversation with the research material. This allowed the author to concurrently and iteratively develop and articulate an understanding of the subject matter both visually and textually. At the end of the research period, the author engaged in a summative analysis [11] of the research material to review the completeness and consistency of the taxonomy.

3.4 Self-Disclosure

While the author has experimented with text-to-image systems and produced digital artworks with these systems, the author is not an artist. The researcher's background is in Computer Science with focus on Human-Computer Interaction (HCI) and Social Computing. The research was conducted not from a technical lens, but a human-centered lens [21]. The author's specific interest in prompt engineering is the text-based interactions of users with text-to-image systems and the novel creative practices that arise from these systems.

4 TAXONOMY OF PROMPT MODIFIERS

Our research points towards there being six different types of prompt modifiers (subject terms, image prompts, style modifiers, quality boosters, repetitions, and magic terms) used by practitioners in the text-to-image art community. We want to highlight that this taxonomy reflects the online community's understanding of prompt modifiers. Technically, a text-to-image system will always produce images of a subject in a certain style and quality, because subjects, style, and quality are inseparably interwoven in the neural network's latent space. The taxonomy presented in this paper reflects the community's understanding of modifiers which was used to categorize modifiers into six different types.

Subject terms indicate the desired subject to the text-to-image system (e.g., "`an old car in a meadow`" or "`a landscape at dawn`"). While it is possible to generate images without subject terms, the subject is essential for controlling the image generation process. On the other hand, since text-to-image systems were trained on images in context of their descriptive text, subject terms can, in some cases, have less control over the outcome. One such case is the artist Zdzisław Beksiński who developed a unique and recognizable style but never provided titles for his artworks. For this reason, early text-to-image systems struggled to reliably reproduce specific subjects in images generated to resemble Beksiński's artworks.

Style modifiers can be added to a prompt to produce images in a certain style. Style modifiers will consistently reproduce a characteristic style. For instance, the modifier "`by Francisco Goya`" will generate digital images in the recognizable style of the late Spanish painter. Other examples of this type of modifier include "`surrealistic painting`", "`oil painting`", "`oil on canvas`"), "`#pixelart`", "`hyperrealistic`", "`Cubism`" or "`cubist`", "`mixed media`", "`cabinet card`", "`in the style of a cartoon`", "`by Claude Lorrain`", or "`in the style of Hudson River School`", to name but a few. As can be seen from the above list, style modifiers can include information about art periods, schools, and styles, but also art materials and media, techniques, and artists. When it comes to the latter, the two modifiers "`by Greg Rutkowski`" and "`by James Gurney`" have become popular in the community of text-to-image art as a means to produce images in a certain style and quality.

Image prompts act similar to subject terms and style modifiers in that they provide the text-to-image system a (visual) target for the synthesis of the image (both in style and subject). Image prompts are typically specified as one or several urls that are added to the textual input prompt or provided in a separate array. Image prompts are different from initial images which were investigated by Qiao et al. [40]. Whereas an image prompt can consist of multiple images, there can only be one initial image. This initial image can be specified as a starting point for the image generation, for instance, for the purpose of enhancing or distorting the initial image. This is made possible because of the iterative nature of the image generation process which typically starts with an image filled with random noise (such as Perlin noise).

Quality boosters can be added to a prompt to increase aesthetic qualities and the level of detail in images. Examples of this type of modifier are the terms "`trending on artstation`" and "`rendered in UnrealEngine`", but also "`highly detailed`, "`masterpiece`", "`AWESOME`", "`#wow`", "`eclectic`, "`fantastic`", "`beautiful`", "`epic`", "`rendering`", or "`rendered in Unreal Engine`", This type of modifier can also be applied in the form of "extra fluff" and verbosity in the prompt which may, in the mind of the practitioner, boost the amount of details and overall quality of the generated image. For instance, the prompt "`painting of an exploding heart`" could potentially be improved by appending the modifiers "`eclectic, detailed, fiery, vfx, rendered in octane, postprocessing, 8k.`"

Repetition can – in the mind of practitioners of text-to-image generation – potentially strengthen the associations formed by the generative system. For instance, the prompt "`space whale. a whale in space`"[2] by @nshepperd1 will likely produce subjectively better results than either of the two subject terms alone. The use of different phrasing and synonyms will cause the text-to-image system to more reliably activate regions in the neural network's latent space that are associated with the subject terms. This is not only an imagined effect. The prompt "a very very very very very beautiful landscape" will, for instance, likely produce a better image than a prompt without repetitions. Technically, this is due to likelihood-maximizing language models becoming stuck in positive feedback loops from repeated phrases [23].

Magic terms introduce randomness to the image that can lead to surprising results. For instance, Twitter user @jd_pressman added the magic term "`control the soul`" to the prompt "`orchestra conductor leading a chorus of sound wave audio waveforms swirling around him on the orchestral stage`"[3]. The term was added to – in his words – produce "more magic, more wizard-ish imagery"[4]. Magic terms thus introduce an element of unpredictability and surprise to the resulting images, often with the intention of increasing the variation in the output. Magic

[2]https://twitter.com/nshepperd1/status/1456584388037148678
[3]https://twitter.com/jd_pressman/status/1457171648293924867
[4]https://twitter.com/jd_pressman/status/1457445367125921793

terms can refer to subjects that are only distantly related to the main subject of the prompt, or they can refer to non-visual qualities, such as the sense of touch (somatosensory), sense of hearing (auditory), sense of smell (olfactory), and sense of taste (gustatory) (e.g., "`feed the soul`" and "`feel the sound`").

In summary, prompt modifiers come in a variety of types and can take different forms. They can, for instance, be added as hash tags (e.g., "`#wow`"), attribution phrases (e.g., "`by [artist]`"), or more complex composite statements (e.g., "`in the style of [artist]`"). Further, not every part of a prompt has the same importance and there are specific affordances of text-to-image systems that are being used in the practice of prompt engineering, as described in the following section.

5 PROMPT WRITING IN PRACTICE

This section provides an overview of how the different types of prompt modifiers are being applied in the practice of prompt engineering with specific focus on the generation of static images from textual and visual input prompts.

While images can be generated from random text or even single characters and emojis [35], the **subject term** is fundamental to the controlled generation of digital images. Consequently, a prompt typically contains at least one subject term. Any other parts of the prompt are optional. It is, for instance, possible to generate artworks with the prompt "car." In practice, however, practitioners use modifiers to improve the resulting images and to exercise more control over the image creation process.

Modifiers are typically added with the intention to either modify the style of the generated image or boost its quality. As mentioned in Section 4, style modifiers and quality boosters do not form a disparate set. Rather, the two types of modifiers can have overlapping effects and the difference between the two types of prompt modifiers is sometimes not fully apparent. The modifier "`Greg Rutkowski`", for instance, exhibits this property. Greg Rutkowski[5] is a contemporary illustrator and concept artist who has been embraced by the text-to-image art community in their practice of prompt engineering. Images generated with the modifiers "`by greg rutkowski`" or "`in the style of greg rutkowski`" are of high quality, texture-rich, and contain a high amount of details. As such, this style modifier is often used as a quality booster in the community, even though a trained eye may tell by the style of the image that the prompt modifier was being used.

Once a style modifier has been added, the style can be reinforced and "solidified" without losing expressivity. **Solidifiers** (in the form of repetitions) can be applied to any of the other types of modifiers (subjects, style modifiers, and quality boosters), although they are most commonly applied to subject terms. Image prompts are a special case in that they can carry both information about the subject and style because of their visual nature. If the textual prompt is aligned with the image prompt, the image prompt can also act as a solidifier. On the other hand, if several images that are different from each other are added to the prompt, the image prompts will contribute to variation in the output. Last, **magic terms** may be optionally added to increase the chance of surprising results. The use of magic terms will result in more variation in the output, while maintaining the overall style.

Each of the six types of prompt modifiers can be assigned **weights**. Weighted terms can be negative to exclude subjects and styles from being generated. For instance, VQGAN-CLIP tends to generate heart-shaped objects with red colour when the prompt contains the word "`love.`" By adding a negative weight to the prompt (e.g., "`heart:-1`"), the system can be instructed not to activate the corresponding latents in its neural network. The resulting images are thus free from heart-shaped objects. Weighted terms can also be used to seamlessly mix styles. For instance, Twitter

[5]https://www.artstation.com/rutkowski

user @c0y0te6 mixed the styles of two artists in the prompt "`a painting of a high prestess [sic] summoning a demon by Ralph McQuarrie:75 | by Zdzislaw Beksinski:25`"[6]. The style of Ralph McQuarrie is, in this case, given precedence over the style of Zdzislaw Beksinski (with a ratio of 3:1).

Table 1 summarizes the iterative nature of prompt writing (c.f. Figure 3). Subject terms are most important for the controlled generation of images and usually written as first step. Modifiers and solidifers are then added to the prompt, either iteratively (image after image) or from learned experience. Last, weights can be applied to exclude or mix subjects and styles.

Table 1. The iterative practice of prompt writing.

Step	Purpose	Prompt modifier	Importance
1	Subject	subject term, initial images	required
2	Modifier	style modifier, quality booster, initial images	optional
3	Solidifier	repetition, initial images	optional
4	Variation	magic terms, initial images	optional
5	Weights	exclusion and mixing	optional

6 DISCUSSION

The availability and accessibility of text-to-image generation as a new computational medium, paired with a specific bundle of technologies and resources that support the ecosystem of this "emerging art scene" [35, 50], have resulted in an explosion of AI generated artworks being shared online. The application of prompt modifiers is key to this emerging creative practice called prompt engineering. Our taxonomy of six different types of prompt modifiers represents an initial work to bringing structure to the creation process and research in text-to-image systems. The taxonomy of prompt modifiers is reified for the sparse HCI literature around prompt engineering as a logical building block in this emerging field.

OpenAI recently announced that the beta of DALL-E 2 would be opened to a million users [31] and Midjourney recently announced their Discord server membership is approaching 1 million users as well [24]. In the near future, everyone will be able to synthesize digital images and artworks from natural language using free or relatively inexpensive means. Gartner estimated that by 2024, 80% of technology products and services will be built by people who are not technology professionals [18]. Increasingly, deep generative models will be used by laypeople without technical expertise and skills and interaction with opaque deep learning models will become increasingly more common in future use cases and applications of artificial intelligence. Therefore, prompt engineering is an emerging research area in the field of Human-Computer Interaction (HCI). However, with the exception of the design guidelines by Liu and Chilton [26] and Qiao et al. [40], there is little scholarly literature in the field of HCI on prompt engineering. Instead, many resources started to emerge from within the online community, such as Smith's Traveler's Guide to the Latent Space [49] and Parsons's DALL-E Prompt Book [36]. The emerging online community around text-to-images generation as well as prompt engineering as a skill and creative practice present numerous opportunities for future research in HCI.

[6]https://twitter.com/c0y0te6/status/1481780797858275329

6.1 Opportunities for Research on Prompt Engineering in HCI

6.1.1 Social aspects of prompt engineering. There are social components to the use of text-to-image generation systems. Prompt engineers face an interesting challenge: Because text-to-image systems were trained on images and text scraped from the Web, users of text-to-image systems need to imagine and predict how other people described and reacted to images posted on the Web. Describing an image in detail is often not enough to achieve optimal results – one has to imagine the image as if it already existed on the Web.

Another social aspect in prompt engineering are the dedicated communities that came into existence only recently. Practitioners of text-to-image art are producing artworks in shared Discord-based chat rooms, such as Midjourney [27]. These dedicated community offer a rich set of social features worth investigating more closely in HCI research. For instance, members on Midjourney have their own profiles that bundle the members' successful creations together with the prompts used to create the images. Recently, Midjourney introduced a 2D map in which members can explore other members based on the similarity of their prompts. Midjourney also has dedicated "group jam" sections in which members can iterate on and further develop other members' works and there is a "theme of the day" section. Long running threads are quite common in this community. Community-learning is an interesting area of research in this regard. How do members receive and seek inspiration in the community? How do novices learn the trade of prompt engineering and is there learning taking place in the community as a whole?

Future work could explore and ethnographically investigate the online community around text-to-image art and its prompt engineering practices in more detail, using the taxonomy presented in this paper as a conceptual starting point or framework.

6.1.2 Human-AI co-creation. While the heart piece of prompt engineering is prompt writing, prompt engineering is only a starting point in some practitioner's creative work flows. Novel creative practices are emerging. For instance, practitioners may develop complex work flows for creating their artworks (e.g., generating initial images with one text-to-image system as a source for inspiration, then continuing on another text-to-image system before finalizing the images in a photo editor). The different affordances of text-to-image systems still need to be reified and systematized in the HCI community. For instance, some text-to-image systems enable the creation of zooming animations, others can complete parts of images (which is called image inpainting [53] and outpainting[7]). These novel creative practices offer a level of interactivity beyond mere generation of static images from textual input prompts. Further, practitioners may make certain idiosyncratic choices when they create text-based generative art (e.g., selecting certain numerical values as seed for the model or adapting the canvas size to certain subject terms). Some of these choices may fall into the realm of folk theories [14, 19] – that is, causal attributions that may or may not be true –, while other choices may be based on the practitioner's experimentation and experience with prompt engineering. Future work could investigate these creative practices, work flows, strategies, and beliefs adopted by practitioners in the text-to-image art community. The emerging research field also offers an opportunity for HCI researchers to make technical contributions [51] in the form of creativity support tools, user interfaces, and interactive experiences to support text-to-image generation, to teach novices the practice of prompt engineering, and to advance the emerging AI generated art ecosystem. Research in this space could make a timely contribution to a novel computational medium and an emerging digital art form.

6.1.3 Bias in image generation systems. Another interesting area for future work is bias encoded in text-to-image generation systems. It has been shown, for instance, that the CLIP model contains

[7] See, for instance, https://twitter.com/adampickard/status/1551584412659335168.

bias[8] and some text-to-image systems prompted with "`princess`" will produce images of women with light skin color, reflecting the bias in the training data toward Western, educated, industrialized, rich and democratic (WEIRD) subjects[9]. OpenAI recently announced that bias was reduced in their DALL-E 2 model [32] (at the cost of potentially reducing signal-to-noise of the generated images[10]).

Responsible deployment of large models and the potential risks are two concerns often listed for not fully releasing a model. While organizations such as OpenAI and Google can be commended for trying to be responsible with their powerful systems, these organizations act paternalistic and impose their value and belief system onto their users which is another source of bias. DALL-E 2, in particular, can be a source of frustration for its users who are often faced with content policy notices for terms relating to war or sexual content (with a threat of account closure if the warning is incurred too often). Pressman et al. recently raised an important point: Humans are sexual beings and the androgynous values imprinted on text-to-image systems with the intent of making them "safe-for-work" deprives users of "a key component of human aesthetic values and experience" [39].

6.1.4 Computational aesthetics and Human-AI alignment. The goal of making computers evaluate and understand aesthetics is much older than text-to-image generation [17]. Recently, there is renewed research on neural image assessment and computational aesthetics. State-of-the-art text-to-image systems increasingly consider human aesthetics in an attempt to produce better images. Prompts are a vast resource for this kind of research, as they encapsulate a person's stated intent. This intent, however, is likely only partially explicit. Research on prompt engineering, therefore, also relates to research on human-AI alignment [16]. This research area is concerned with teaching artificial intelligence to understand human values. Prompts for text-to-image generation systems could form an interesting study resource for this kind of research.

Besides the above opportunities for research on prompt engineering in HCI, there are also broader interdisciplinary implications of ethical and societal relevance, as outlined in the following section.

6.2 Broader Implications for Human-AI Interaction

Research on prompt engineering has broader implications and is not only limited to the field of text-to-image synthesis and AI generated art, but also relevant to the interaction of humans with deep learning models and artificial intelligence in general.

6.2.1 AI and the future of creative work. There is much potential for deep learning to disrupt and transform entire sectors of the creative economy. Recently, there has been an interest into developing generative systems that are able to synthesize more complex outcomes. For instance, Hong et al. presented CogVideo, a system for text-to-video generation [25]. In the future, we may see deep generative models with generative capabilities that transcend what we can imagine today. Deep generative models could, for instance, create entire interactive story-driven worlds and games from short text prompts.

Such powerful AI-based systems will have implications for the future of creative work. Artificial intelligence will not only transform the way we interact with computers and perform work online, but also the content of our work and the human agency in the work. An example of an application that has such transformative potential is OpenAI's Codex [3, 52]. Codex is a large language model that interprets commands in natural language and generates programming code. In the future, instead of typing code, we will be able to describe a software and its expected outputs in natural language. Deep learning models, such as Codex, BLOOM [1] or other large "foundation models" [2],

[8]See https://twitter.com/RiversHaveWings/status/1432100170645180416.
[9]See https://twitter.com/EMostaque/status/1495323912951021568.
[10]See https://twitter.com/minimaxir/status/1549070583035416576.

will then generate executable software code based on the human's spoken or written input prompts. This change in the agency of humans and computers could be transformative to creative work, such as software development. Similar technology has already found application in GitHub's CoPilot[11], an "AI pair programmer" assisting its users in auto-completing programming code.

Low-code and no-code tools for creating online products and experiences will become increasingly common in the future and declarative machine learning systems may – as a next wave of machine learning – bring machine learning to non-coders [28]. This technology will extend the currently rather narrow focus of prompt engineering on language models and text-to-image synthesis to more broader application domains.

6.2.2 Beyond text-to-image generation. The use case of text-based generative art discussed in this paper is but one of many application areas of prompt engineering, with implications for the future of creative work and Human-AI Interaction (HAI) in general. The latter can be viewed from many different perspectives, such as human-centered AI [48], human-AI partnerships [42], and human-AI cooperation [4]. Irrespective of the term used to describe our relationship with AI, we will increasingly interact with opaque models through natural language in the future.

Research on how input prompts can be effectively formulated is therefore timely and important. Research in this space will advance our understanding of how people can effectively interact with machine learning models. The research will inform the design of future applications and systems powered by Artificial Intelligence (AI) and advance the understanding of the needs of different stakeholders using these AI-driven systems. A better understanding of users' interactions with AI could not only advance research into interpretability and explainability of AI-driven systems, but also provide valuable insights into how web-based systems can be designed for fairness, accountability, and transparency. These three constituents could be a step stone to support the formation of trusting social relationships with artificial intelligence.

7 CONCLUSION

This paper provided an investigation of prompt engineering practices and contributed a taxonomy of six different types of prompt modifiers used in the community. Style and quality modifiers are the two most prominent and commonly used types of modifiers. The taxonomy of prompt modifiers lays the foundation for future structured investigations into prompt engineering for text-to-image generation and AI generated art. We further demonstrated how these prompt modifiers may be applied to make image generations more reliable or variable. The paper discussed several opportunities for future research in the field of HCI on the fast-moving text-to-image community and touched on the broader implications for creative work and interaction of humans with artificial intelligence beyond the use case of text-to-image generation.

REFERENCES

[1] BigScience Initiative. 2022. Introducing The World's Largest Open Multilingual Language Model: BLOOM. https://bigscience.huggingface.co/blog/bloom

[2] 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, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li,

[11] copilot.github.com

Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan 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 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:2108.07258 [cs.LG]
[3] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs.LG]
[4] Jacob W. Crandall, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A. Goodrich, and Iyad Rahwan. 2018. Cooperating with machines. *Nature Communications* 9, 1 (2018), 12 pages. https://doi.org/10.1038/s41467-017-02597-8
[5] Lyall Crawford. 1996. Personal ethnography. *Communication Monographs* 63, 2 (1996), 158–170. https://doi.org/10.1080/03637759609376384
[6] Katherine Crowson. 2021. CLIP Guided Diffusion HQ 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj
[7] Katherine Crowson. 2021. VQGAN+CLIP(Updated).ipynb. https://colab.research.google.com/github/justinjohn0306/VQGAN-CLIP/blob/main/VQGAN%2BCLIP(Updated).ipynb
[8] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. https://doi.org/10.48550/ARXIV.2204.08583
[9] Norman K. Denzin and Yvonna S. Lincoln. 2017. *The SAGE Handbook of Qualitative Research* (5th ed.). SAGE, Thousand Oaks, CA.
[10] Disco Diffusion Contributors. 2022. Disco Diffusion. https://github.com/alembics/disco-diffusion
[11] Margot Duncan. 2004. Autoethnography: Critical Appreciation of an Emerging Art. *International Journal of Qualitative Methods* 3, 4 (2004), 28–39. https://doi.org/10.1177/160940690400300403
[12] Remi Durant. 2021. Artist Studies by @remi_durant. https://remidurant.com/artists/
[13] Carolyn Ellis, Tony E. Adams, and Arthur P. Bochner. 2011. Autoethnography: An Overview. *Historical Social Research / Historische Sozialforschung* 36, 4 (138) (2011), 273–290. http://www.jstor.org/stable/23032294
[14] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, 2371–2382. https://doi.org/10.1145/2858036.2858494
[15] Harmeet Gabha. 2022. Disco Diffusion 70+ Artist Studies. https://weirdwonderfulai.art/resources/disco-diffusion-70-plus-artist-studies/
[16] 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
[17] Philip Galanter. 2012. *Computational Aesthetic Evaluation: Past and Future.* Springer Berlin Heidelberg, Berlin, Heidelberg, 255–293. https://doi.org/10.1007/978-3-642-31727-9_10
[18] Gartner. 2021. Gartner Says the Majority of Technology Products and Services Will Be Built by Professionals Outside of IT by 2024. Press release. https://www.gartner.com/en/newsroom/press-releases/2021-06-10-gartner-says-the-majority-of-technology-products-and-services-will-be-built-by-professionals-outside-of-it-by-2024
[19] Susan A. Gelman and Cristine H. Legare. 2011. Concepts and folk theories. *Annual Review of Anthropology* 40 (2011), 379–398. https://doi.org/10.1146/annurev-anthro-081309-145822
[20] Raymond L. Gold. 1958. Roles in Sociological Field Observations. *Social Forces* 36, 3 (1958), 217–223. http://www.jstor.org/stable/2573808
[21] Mark Guzdial. 2013. Human-Centered Computing: A New Degree for Licklider's World. *Commun. ACM* 56, 5 (may 2013), 32–34. https://doi.org/10.1145/2447976.2447987

[22] 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

[23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR '20)*. 16 pages.

[24] David Holz. 2022. Hey everyone, this server is getting BIG! Announcement on Midjourney. https://discord.com/channels/662267976984297473/952771221915840552/1003404579897102387

[25] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. https://doi.org/10.48550/ARXIV.2205.15868

[26] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 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

[27] Midjourney. n.d.. Midjourney.com. https://www.midjourney.com

[28] Piero Molino and Christopher Ré. 2021. Declarative Machine Learning Systems. *Commun. ACM* 65, 1 (dec 2021), 42–49. https://doi.org/10.1145/3475167

[29] Carman Neustaedter and Phoebe Sengers. 2012. Autobiographical Design in HCI Research: Designing and Learning through Use-It-Yourself. In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. Association for Computing Machinery, New York, NY, 514–523. https://doi.org/10.1145/2317956.2318034

[30] Anna Notaro. 2020. State-of-the-art: AI through the (artificial) Artist's Eye. In *Proceedings of EVA London 2020*. BCS, The Chartered Institute for IT, London, UK, 322–328. https://doi.org/10.14236/ewic/EVA2020.58

[31] OpenAI. 2022. DALL·E Now Available in Beta. https://openai.com/blog/dall-e-now-available-in-beta/

[32] OpenAI. 2022. Reducing Bias and Improving Safety in DALL·E 2. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/

[33] OpenAI. nd.. Completion – OpenAI API. https://beta.openai.com/docs/guides/completion

[34] OpenAI. nd.. Examples – OpenAI API. https://beta.openai.com/examples

[35] Jonas Oppenlaender. 2022. The Creativity of Text-based Generative Art. , 9 pages. https://doi.org/10.48550/ARXIV.2206.02904

[36] Guy Parsons. 2022. The DALL·E 2 Prompt Book. https://dallery.gallery/wp-content/uploads/2022/07/The-DALL%C2%B7E-2-prompt-book-v1.01.pdf

[37] Sarah Pink, Heather Horst, John Postill, Larissa Hjorth, Tania Lewis, and Jo Tacchi. 2016. *Digital Ethnography: Principles and Practice*. SAGE, London, UK.

[38] John Postill and Sarah Pink. 2012. Social Media Ethnography: The Digital Researcher in a Messy Web. *Media International Australia* 145, 1 (2012), 123–134. https://doi.org/10.1177/1329878X1214500114

[39] John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. 2022. *Simulacra Aesthetic Captions*. Technical Report Version 1.0. Stability AI. url https://github.com/JD-P/simulacra-aesthetic-captions .

[40] 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

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[42] Sarvapali D. Ramchurn, Sebastian Stein, and Nicholas R. Jennings. 2021. Trustworthy human-AI partnerships. *iScience* 24, 8 (2021), 13 pages. https://doi.org/10.1016/j.isci.2021.102891

[43] 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

[44] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[46] Robin Rombach, Andreas Blattmann, and Björn Ommer. 2022. Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models. https://doi.org/10.48550/ARXIV.2207.13038

[47] 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

[48] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[49] 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

[50] Charlie Snell. 2021. Alien Dreams: An Emerging Art Scene. https://ml.berkeley.edu/blog/posts/clip-art/

[51] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (2016), 38–44. https://doi.org/10.1145/2907069

[52] Wojciech Zaremba and Greg Brockman. 2021. OpenAI Codex. https://openai.com/blog/openai-codex

[53] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. 2020. *Text-Guided Neural Image Inpainting*. Association for Computing Machinery, New York, NY, 1302–1310. https://doi.org/10.1145/3394171.3414017