Exploring user interactions with Generative AI for Product Design

Allison (I-Ping), Lo
1465015

August 3, 2023

Utrecht University

Department of Computing Sciences

Computing Science

Exploring user interactions with Generative AI for Product Design

Allison (I-Ping), Lo
1465015

1. Reviewer Ioanna Lykourentzou
Department of Computing Sciences
Utrecht University

2. Reviewer Johan Jeuring
Department of Computing Sciences
Utrecht University

Thesis Supervisor Ioanna Lykourentzou

Day-to-Day Supervisor Ioanna Lykourentzou

August 3, 2023

Allison (I-Ping), Lo
Exploring user interactions with Generative AI for Product Design
Computing Science, August 3, 2023
Reviewers: Ioanna Lykourentzou and Johan Jeuring
Thesis Supervisor: Ioanna Lykourentzou
Day-to-Day Supervisor: Ioanna Lykourentzou

Utrecht University
Department of Computing Sciences
Heidelberglaan 8
Postbus 80125, 3508 TC Utrecht

Abstract

This project aims to investigate the potential of Generative AI for product design by exploring the application of a prompt-to-image generative AI tool(which is based on the stable diffusion model) in generating new products (in this case: bikes) tailored to individual users' specific needs and preferences. Traditional recommender systems select and recommend an item from a pre-existing database, while Generative AI generates new items based on user input. The research questions that this thesis will focus on are: 1) "How are users currently using a prompt-to-image Generative AI tool to design products that meet their requirements?" and 2) "What are the approaches or prompts that can help to generate better results?" The study aims to observe and extract the way of prompting or action patterns in how users utilize an existing prompt-to-image generative AI tool and see what help to generate a better or preferable design. Understanding the features and functionality essential for designing their ideal product to align future tool design is another goal of this research.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Recommender Systems 3
2.2 Generative AI 4
2.2.1 Generative adversarial networks (GANs) 4
2.2.2 Diffusion model 5
2.3 Generative AI for product design 6
2.4 Off-shelf prompt-to-image softwares using Generative AI 10
2.4.1 DALL·E 11
2.4.2 Midjourney 11
2.4.3 Artbreeder 12

2.4.4 Leonardo.AI 12

3 Methodology **15**
3.1 Participants 15
3.2 Execution of the experiment 16
3.2.1 Pre-Task tutorial 16
3.2.2 Task Accomplishment 16
3.2.3 Post-Task Interviews 17
3.2.4 Crowd-sourced evaluations 18
3.3 Data Analysis 18
3.3.1 The Quantitative Data 20
3.3.2 The Qualitative Data 22

4 Results **27**
4.1 Exploring User Approaches in Designing Products with Prompt-to-Image Generative AI Tools 27
4.1.1 Users start with global prompts and move to local refinements, spend most of their time on local 27
4.1.2 On global, users focus on novelty, while in the local mode, they focus on feasibility 28
4.1.3 User write lengthier prompts in Global Editing Phases and shorter prompts in Local editing phase 31
4.1.4 Shared and Synergistic prompting between Feasibility and Aesthetics 32
4.1.5 Common topics in people's prompts 33
4.2 Strategy of success for the Prompt-to-Image Generative AI Tools . . . 34
4.2.1 Allowing the AI to think for you? Also remember to direct it toward your specific creative design 34
4.2.2 Multi- vs. Mono-Criteria Prompts: Broad Start is more Beneficial 36
4.2.3 To Boost novelty, try more novel-related prompts, but more feasibility-related prompts do not guarantee better feasibility 37
4.2.4 The Power of Aesthetic Prompts to Enhance Feasibility (as well as Aesthetics) 38
4.2.5 The More Is Not Always the Merrier: Lengthy Prompts Aren't Always the Key to Success 39
4.2.6 Decoding the Keywords of Successful Prompting 40
4.2.7 Be aware of the Trade-Off: Relationship Between Feasibility and Novelty 42

5 Discussion, Limitations and Further Work **45**

5.1 Design Recommendations 45
5.2 Limitations 47
5.3 Future Work 48

6 Conclusion **51**

Bibliography **55**

1 Introduction

Determining the optimal design of a product is essential for a wide range of consumers and applications, from identifying the most suitable coffee mug to the most appropriate interior design for one's home, or from an ideal bicycle to an automobile. Likewise, understanding the ideal product for customers is crucial for businesses as it enables them to design and manufacture products that align with consumer preferences and demands.

Currently, recommender systems are widely employed to assist customers in identifying appropriate products. The field of recommender systems research has been a significant area of study in recent decades; the research in this field continues to push the boundaries of the capabilities and performance of recommender systems, making them increasingly effective in providing personalized and relevant recommendations to users.

In 2020, the discussion of Generative AI in various applications has emerged. These models have gained popularity for their ability to "generate" new images and other forms of data. In this paper, we aim to investigate the potential of Generative AI for product design by exploring the application of Generative AI in generating "new" products tailored to individual users' specific needs and preferences. The question that motivates this study is: "What if instead of recommending products from a large but fixed product dataset, a system could help users design their ideal product, something totally tailored to their needs but still feasible to produce?"

Recommender systems based on Generative AI are a relatively nascent area of research. The primary distinction between Generative AI-based and traditional recommender systems is that, instead of selecting and recommending an item from a pre-existing database, Generative AI generates new items based on user input. However, there are limitations in the current state of Generative AI-based recommender systems, particularly in their application to text input and their ability to produce feasible products. For example, Generative AI may generate an interior design suggestion that is impossible to implement due to the laws of physics, thus underlining the need for further research and development to address these limitations and improve the effectiveness and feasibility of Generative AI-based recommender systems.

The research questions that we will focus on in this thesis are therefore:

1. "How are users currently using prompt-to-image Generative AI tools to design products that meet their requirements?" To find out the answer, we will observe how users use a currently existing system that generates product designs based on Stable Diffusion technology. The complete setup of observation will be illustrated in a later section.

2. Identifying the essential parameters and flows that users consider important for designing their ideal product in the previous research question, our next question is: "What prompts and approaches yield better results?" By understanding what prompts and approaches are effective, we can provide suggestions about optimizing the design process and enhancing the overall user experience.

2 Related Work

2.1 Recommender Systems

A recommender system is a type of system that uses data-processing techniques to recommend items to users. The goal of a recommender system is to provide personalized and relevant recommendations to users based on their preferences, interests, and past behavior.

The history of recommender systems can trace back to the early days of information retrieval and collaborative filtering. One of the earliest forms of recommendation systems was the implementation of collaborative filtering, which began in the early 1990s [Pen+13]. It described a system that used a combination of demographic information and personality diagnosis to make recommendations to users.

In the late 1990s and early 2000s, a number of other companies, such as Amazon [LSY03] and Netflix [GH16], also began to use collaborative filtering to make recommendations to users.

Nowadays, recommender systems can be used in a wide range of applications in our daily life, such as e-commerce, social media, music and video streaming, and even more. They can recommend products, services, or content to users based on their past behavior, such as browsing history, purchase history, or search queries.

As the Recommender System Handbook [RRS15] indicated, there are several different types of traditional recommender systems, such as:

1. Content-based filtering: which makes recommendations based on the characteristics of the items that a user has previously interacted with.
2. Collaborative filtering: which makes recommendations based on the preferences of similar users.
3. knowledge-based: enabling the generation of advice based on explicit human knowledge about the item assortment, user preferences, and recommendation criteria (i.e., which item should be recommended in which context)
4. Hybrid recommender systems: which combine the above methods.

With the advent of big data and the increasing availability of large amounts of user data, the field of recommendation systems has grown rapidly. In recent years, new methods such as matrix factorization [KBV09], which models the user-item interactions as a low-dimensional latent space, and deep learning [Zha+19], which uses neural networks to model the user-item interactions. have been used to improve the precision and diversity of recommendations.

However, there are several limitations and drawbacks of traditional recommender Systems, such as : Difficulties of scalability, sparsity, lack of diversity, and difficulties of personalizing. As traditional RS requires massive data on users and items to make accurate recommendations, it will be computationally expensive and inefficient whether algorithms try to map the features or attributes between users or items. Imagine how much data space will be taken by just adding an item or a user, or how much data space and calculation time are wasted on the rarely used or less informative items/users.

Another limitation crucial to this thesis is that typical recommender systems **are limited to their/pick from a specific database**, so the product must already exist to recommend it. They are not suitable for custom-made designs or for generating new products for customers, and it leads to the problem of lack of diversity and difficulties of personalization.

2.2 Generative AI

Generative AI is a rapidly evolving field that involves the creation of algorithms and models that can generate novel content in various domains, such as images, text, and music. Its primary goal is to imitate the intricate creative process by leveraging existing datasets in order to identify underlying patterns and generate outputs that closely resemble the characteristics of the training examples. Within this domain, two influential techniques have emerged: Generative Adversarial Networks (GANs) and the diffusion model.

2.2.1 Generative adversarial networks (GANs)

Generative adversarial networks (GANs) are a relatively new technique in the field of machine learning presented by [Goo+20] in 2014. The GANs are built by a pair of neural networks and training them under the idea that one network's gain is another network's loss. One network, often called the "generator model," will be trained to

generate new examples. The other, often known as the "discriminator model," will try to distinguish whether the examples are real data or generated data.

Generative Adversarial Networks (GANs) have emerged as a promising approach to address the long-standing problem of traditional machine learning, which often requires a vast amount of data to train and fine-tune models [Kar+17]. Unlike traditional machine learning algorithms, which rely on supervised or unsupervised learning techniques, GANs utilize a generative model trained to produce new data samples similar to those in the training data [RMC15].

Although Generative Adversarial Networks (GANs) have shown considerable promise, training and fine-tuning these models can be a challenging task. Mode collapse, a common issue with GANs, describes a situation when the generator only produces a limited selection of outputs that are simple for the discriminator to recognize. As a result, generated data samples might be of poor quality, which could harm the performance of the model overall. Several methods are suggested to address this problem. Such as injecting noise into the generator's input [Sal+16], using diverse loss functions [Mao+17], and applying regularization methods [Miy+18], to promote the diversity of the generated data and improve the overall quality of the model outputs.

After years of ongoing research and refinement, the practical applications of GANs have expanded significantly. Within the design domain, GANs can serve as effective tools for generating images, 3D objects, and videos from textual descriptions or other visual inputs. For example, in paper [Ree+16], they demonstrate the capability of generating images from text descriptions; ; in this research [Iso+17], their model can transfer styles or features from image to image; and in this paper, [VPT16], they successfully generate videos with scene dynamics.

2.2.2 Diffusion model

The diffusion model was introduced by Sohl-Dickstein et al in 2015, which laid the groundwork [ND21] proposed the diffusion process as an alternative paradigm to traditional generative models.

Extending from this base, Dinh et al. [DSB16] introduced a flow-based generative model employing invertible transformations. This extension broadened the scope of diffusion models and their potential applications.

Following by a paper that proposed a generative model called FFJORD (Free-form Jacobian of Reversible Dynamics) [Gra+18], they introduced a continuous-time

diffusion process capable of generating high-quality samples and enabling efficient inference.

Kingma and Dhariwal advanced diffusion models [KD18] that combined flow-based models with invertible 1x1 convolutions. This architectural innovation improved the expressive power and computational efficiency of diffusion models.

Ho et al.'s paper [Ho+19] further improved flow-based generative models. Their research focused on enhancing the quality and diversity of generated samples.

Diffusion models offer unique advantages compared to GANs. They can introduce conditioning on multiple features, guaranteeing a more fine-grained control over the generated images and allowing data quality and diversity manipulation [GL23].

The training process of Diffusion models is more stable, avoiding mode collapse. A paper by OpenAI researchers [DN21] has indicated that diffusion models can achieve image sample quality superior to the GANs models.

However, it is worth mentioning the main drawbacks of diffusion models, which are that they require longer training times and are computationally intensive. This is because of their inherent complexity and the sequential nature of the diffusion process. During training, multiple steps of noise conditioning need to be applied, which can be quite computationally demanding. Also, diffusion models often have more parameters and intricate architectures than traditional generative models, increasing the computational requirements and extending the training time.

2.3 Generative AI for product design

As the field of AI-assisted tools is still emerging, research papers in the area of users' interactions with AI-assisted tools remain relatively sparse.

GANSpace [Här+20] is a technique for discovering interpretable controls for Generative Adversarial Networks (GANs) by performing Principal Component Analysis (PCA) in activation space. In their tool, as fig 2.1 shows, they provide more than 20 slider controls for interactive exploration of the principal directions, and layer-wise application is enabled by specifying a start and end layer for which the edits are to be applied. Another paper [KG22] also lets users control the sliders to adjust the parameters, including blending of style suggested by the generative models. The UI is shown in fig 2.2

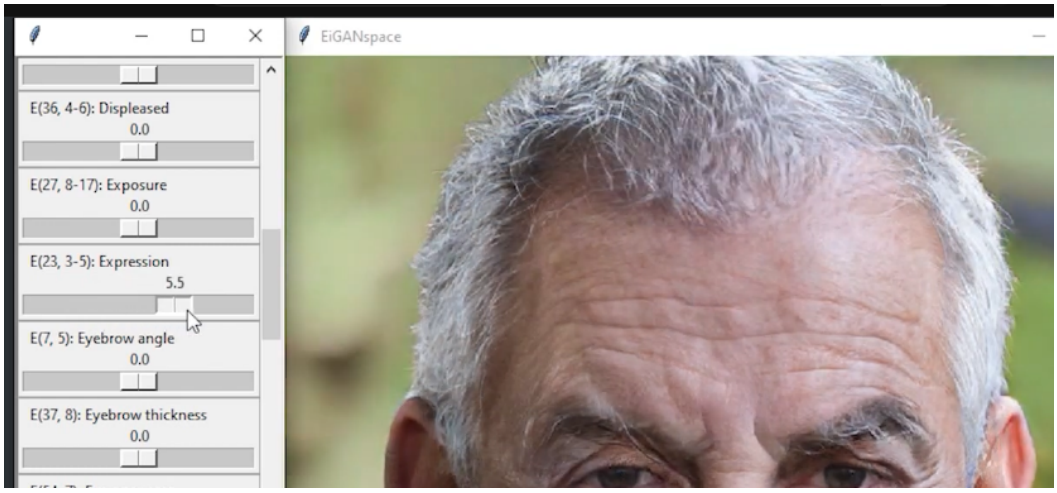


Fig. 2.1: Multiple sliders used in UI of the system in the GANSpace paper [Här+20]

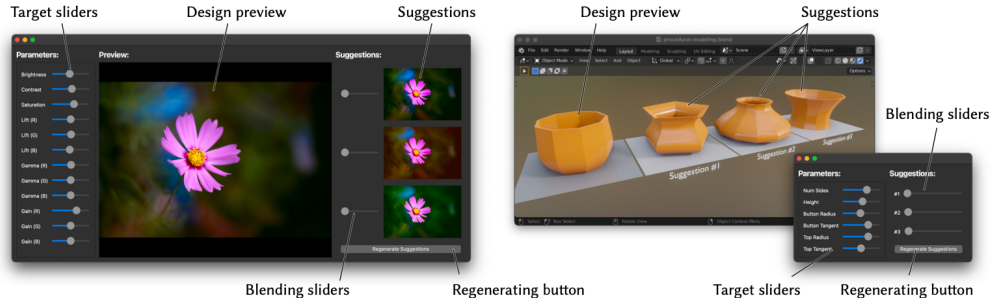


Fig. 2.2: UI of the system in the BO as assistant paper [KG22]

Sliders are commonly used in editing tools that let users control elements to adjust certain aspects, such as brightness, color balance, or even the level of application of a particular style. Previous research [DMB22] held an experiment to see if more sliders lead to faster or more accurate task performance. However, the results state that more control dimensions (sliders) significantly increase task difficulty and user actions.

In another study [ZB21], they developed an interactive interface that facilitates GAN exploration by providing users with a grid-like view of sampled images, thus allowing them to navigate the latent space. Their UI components are shown in fig 2.3, including a tool palette(zoom in, zoom out, zoom into region, pivot, snapshot, randomize, and undo and redo), a current working image gallery (with 25 GAN-generated images organized in a 5 $\times$ 5 grid), gallery snapshots(saving the generated images history of 5 * 5 grid), and user-selected quality images for their interactive interface for GAN exploration. The users are allowed to click and select

an image that matches their object.

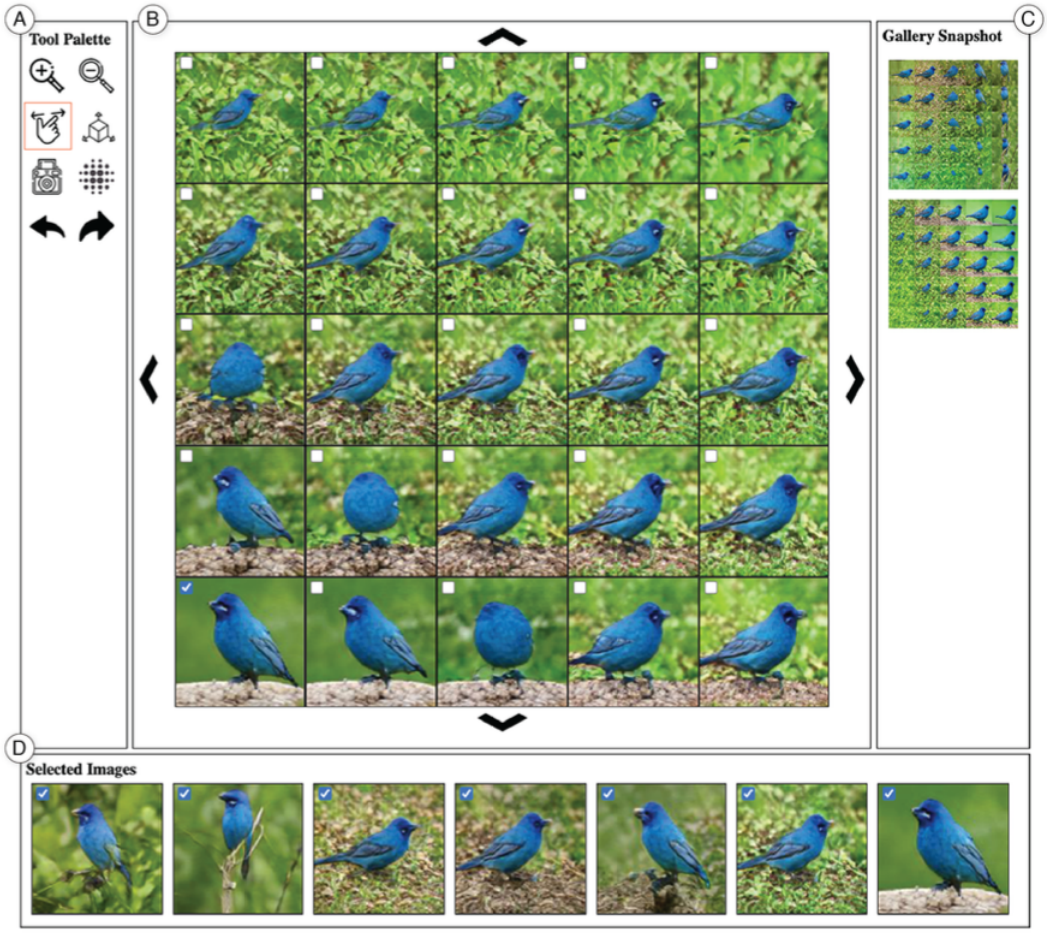


Fig. 2.3: Image gallery UI with 25 GANs-generated images organized in a 5 × 5 grid in the automatically generated image galleries paper [ZB21]

Sketch-based design tools like Forte[Che+18], and DreamSketch[Kaz+17], provide drawing canvas interfaces that transform sketching to 2D or even 3D designs with loading scenarios. Other controlling toolkits, like sliders, enable users to adjust the amount of material for generating the structures or control how the generated results are similar to the original sketch. Screenshots of their UI are shown in fig 2.4 and fig 2.5. The user reviews in the paper got positive feedback that the capability of generating structures from sketches is interesting, and it also makes the idea of topology optimization accessible to people.

Image editing or generating by language-based input is also commonly seen in AI-assisted tools. For example, DALL·E [Ram+21] allows the user to use natural

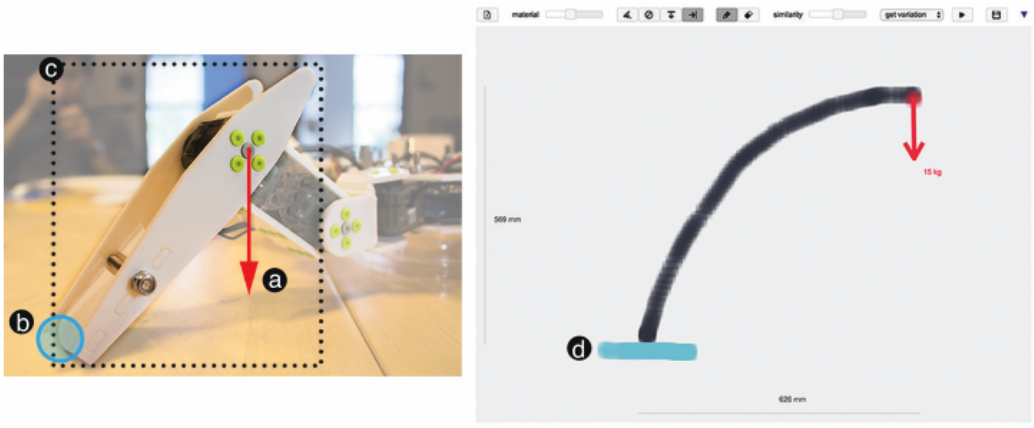


Fig. 2.4: The UI of Forte includes a canvas feature that allows users to sketch out the items and adjust with provided toolkit [Che+18]

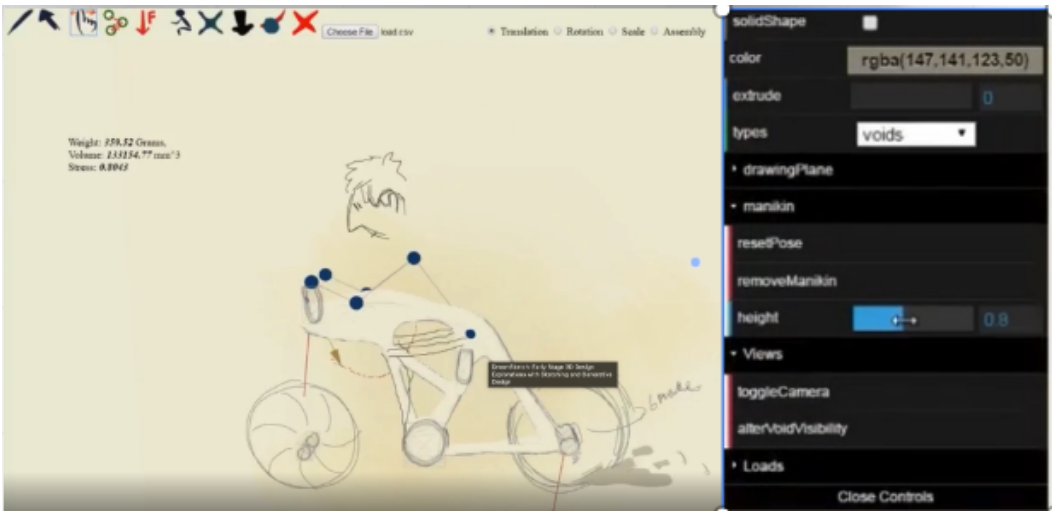


Fig. 2.5: The UI of DreamSketch includes a canvas feature that allows users to sketch out the items and adjust with provided toolkit [Kaz+17]

language textual input to interact with the back end AI model, such as "Create an image of a cat in the fashion of Van Gogh. or text prompt like "imagine jasmine in the wildflower –w 600 –h 300". A longer paragraph or more specific text input can also improve the image it generates. Another tool people often compare with DALL·E is Midjourney, which can mainly be accessed via a Discord bot on their official Discord channel. Users can interact with the bot either through direct messaging or by inviting it to a third-party server. To create images, users need to enter a prompt after typing in the '/imagine' command, and the bot will generate an image in response. As these tools are the most well-known and comparably mature products in the industry, they will be used as baseline tools in the experiment of this paper. A detailed explanation of their functionality and user interface will therefore be given

in the following chapter.

As for the production aspect, designs that are aesthetically pleasing and conform to certain design constraints [YM20] are not always enough. Physics simulation capabilities are also crucial functions to be included in the tools. This might be a complex problem in the past, but research lately [Shu+20] has proved that it might be manageable.

Some people also use multiple AI-assisted tools sequentially. For example, as ChatGPT is known for its powerful ability to understand and generate natural language content, a designer can first create a detailed user persona and ask ChatGPT [Ouy+22] to design a bag, and after getting the textual answer, use it as the input for Midjouney or DALL·E. It could help you generate the image of the design. With the powerful tools to explore the latent space and refine it by either more textual input or provided tools, after that, if needed, export the image and feed to another tool that can transfer 2D images to 3D items, the result might be already good enough for actual production use. This highlights the fact that no single AI tool currently exists that can fully satisfy the diverse needs of all users. Addressing this issue and suggesting improvement upon existing tools will be a key focus of this research.

2.4 Off-shelf prompt-to-image softwares using Generative AI

There is a wide range of tools that are based on generative AI models, which can utilize text, images, or a combination of both to generate image designs. Some popular options include GANs-based tools like Artbreeder and diffusion model-based tools like DALL·E, Midjourney, and Leonardo.AI.

As all these tools allow users to input text and receive corresponding images, their use cases and user flows exhibit certain similarities. However, notable distinctions arise in terms of UI layout. As some tools offer more features within the editing canvas and image fine-tuning section, the variations grant users greater freedom to manipulate and refine the generated results according to their preferences. Each

tool has its own strengths and limitations, which will be discussed in subsequent sections.

2.4.1 DALL·E

As shown in Figure 2.6, the process of generating images with DALL·E involves simply entering a text command or "prompt" into the provided input field and clicking "Generate". After a short processing time, four images are generated, each of which can be clicked to view more details. From this, users can create variations or make edits using the built-in editing function. This includes selecting and marking specific areas of the image that DALL·E can then modify using a new text prompt. Simply describe the desired changes in the new prompt, and DALL·E will apply them accordingly. Additionally, DALL·E can extend the image based on the selected area using the prompt's instructions. Users can keep track of each batch of generated images using the history section.

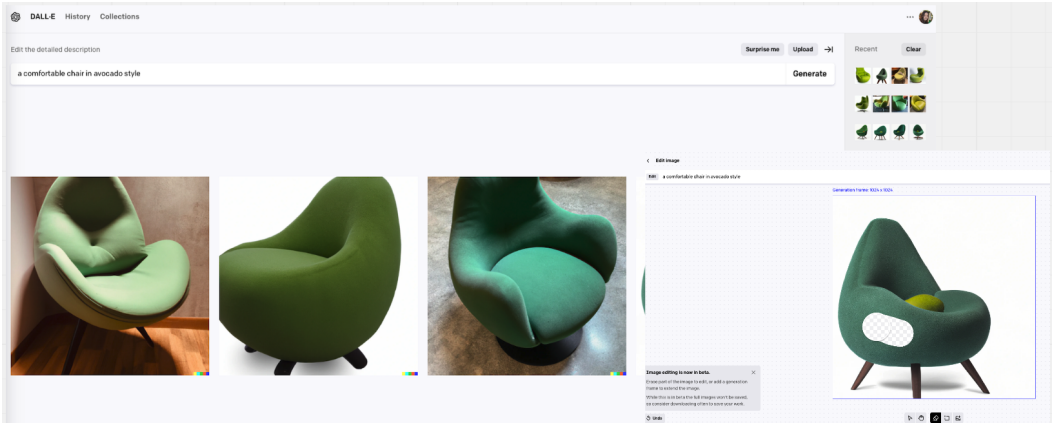


Fig. 2.6: UI of DALL·E. A text input box, four generated images, and a history section make up the main components. Highlights of the canvas toolkits include the extend and erase functions, which give you more opportunities to thoroughly edit the image. Source: DALL·E

2.4.2 Midjourney

As demonstrated by 2.7, Midjourney is now a chatbot available on the Midjourney Discord server. To begin generating images, the user starts by typing "/imagine prompt" in the chat box and entering a description of the image the user wishes to create in the "prompt" field. After the user has entered the prompt, press the "return" key to submit your message. The Midjourney Bot will take approximately

one minute to generate four image options with two rows of buttons. The "U" buttons will upscale the selected image, generating a larger version with additional details. The "V" buttons will create slight variations of the selected image, generating a new image grid that shares the overall style and composition of the original image. And the "re-roll" button will rerun the original prompt, producing a new grid of images based on the same input.

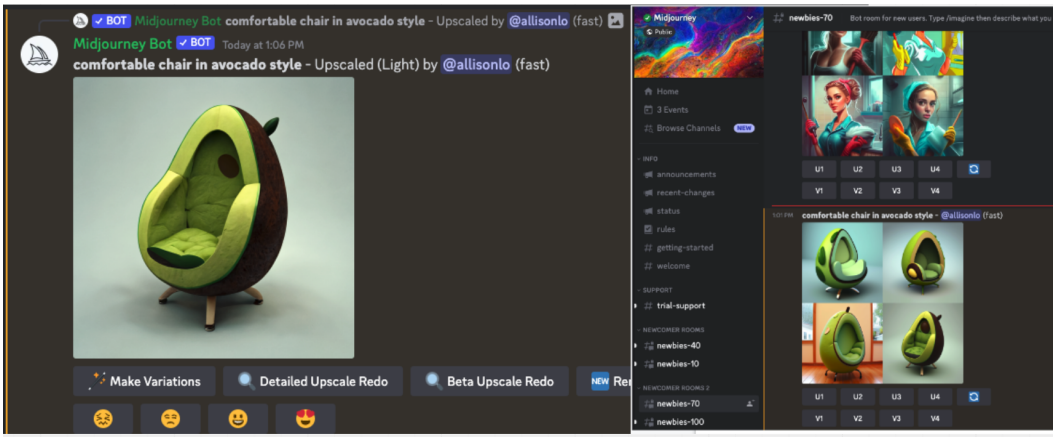


Fig. 2.7: UI of Midjourney. Although it has considerably less manipulable functionality than other tools, the aesthetic quality of the image Midjourney generated is believed to be better than that of other tools. Source: Midjourney

2.4.3 Artbreeder

As illustrated in fig 2.8, the Collage tool on Artbreeder enables users to create a distinctive image by mixing shapes and images from a library or drawing their own with a text prompt describing their vision for the collage. Artbreeder generates the image based on the input. In addition, users have the ability to adjust the AI render intensity of the model to fine-tune the outcome and make it closer or farther away from the original input. This feature allows greater customization and control over the final product.

2.4.4 Leonardo.AI

As shown in the 2.9, Leonardo.AI offers functionality similar to DALL·E, such as text-to-image conversion with optional image input as a base image and a canvas for editing images. It is a free tool that allows users to use the popular open-sourced diffusion model: Stable Diffusion 2.1. Moreover, it also provides users with

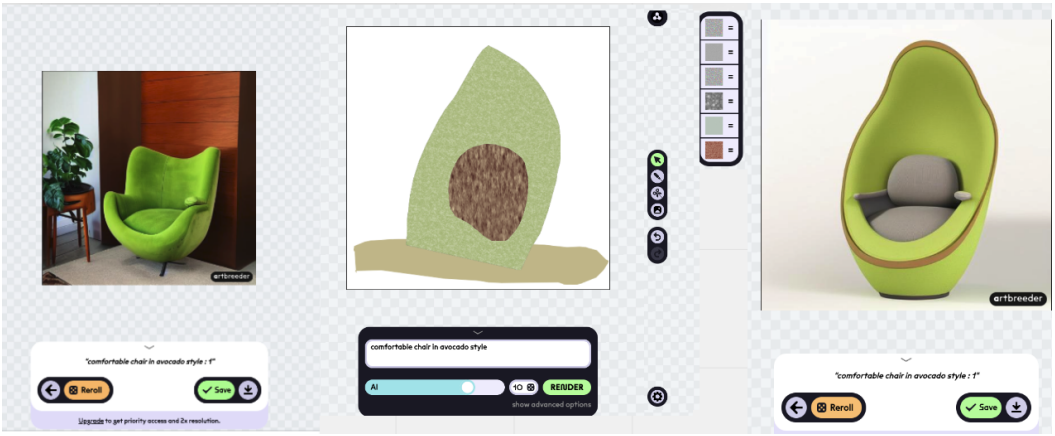


Fig. 2.8: UI of Artbreeder. It is one of a kind to have the option to use text, drawing, and existing shape as multiple inputs and to allow the user to adjust the weight of the text and graphical input of the model. Source: Artbreeder

additional model controls, including the Guidance Scale feature, which enables users to adjust the weight of the prompt and determine how many steps count to render. As a result, users can achieve more detailed and intricate images by increasing the number of steps, although it may take longer to complete. This added level of control provides greater flexibility and customization options for users when generating images.

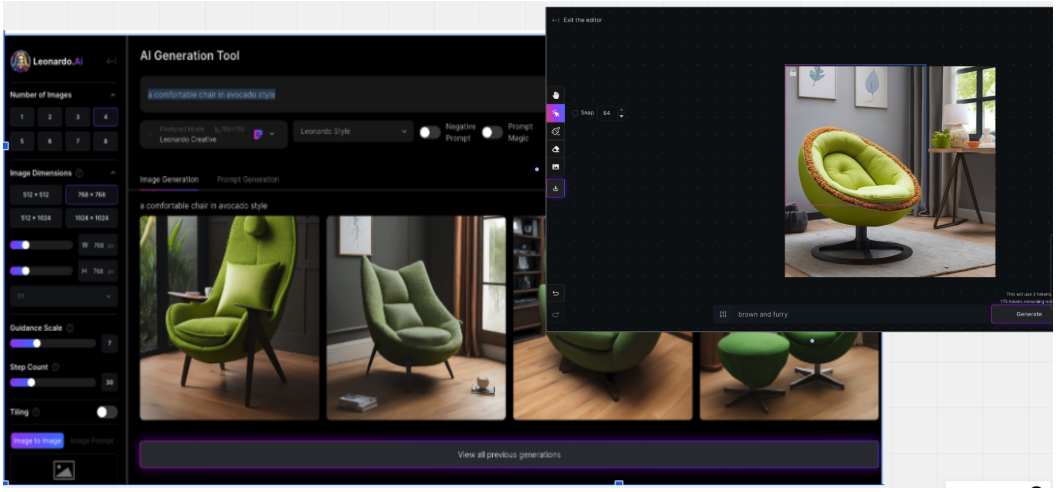


Fig. 2.9: UI of Leonardo.AI. As a free tool, it offers users a remarkable level of freedom to adjust the generated model and image. Furthermore, the algorithm behind this tool, Stable Diffusion, is highly dependable. Source: Leonardo.AI

3 Methodology

This chapter provides an in-depth discussion of the methodology employed in the study. The methodology encompasses various aspects, such as participant selection, the execution of the experiment, and data analysis.

We adopted the think-aloud protocol to capture a comprehensive understanding of user behavior and insights into their utilization of the generated AI tool for exploring the generative design space and creating images based on specific criteria. The entire process was recorded and transcribed, facilitating the observation and analysis of user actions.

Each participant was given a tool called Leonardo.AI, which was introduced in a preceding section on off-the-shelf software. This tool encompasses a wide range of features, offering users an extensive platform for experimentation.

Upon completing the assigned task, participants underwent post-task interviews and were provided with a brief questionnaire. These additional measures aimed to gather further qualitative data. Subsequently, all the collected data were subjected to thorough analysis.

In order to obtain a broader range of ratings on the quality of the generated images in terms of feasibility, novelty, and aesthetics, we enlisted the assistance of crowd-sourced evaluations as it can help to recruit diverse participants.

3.1 Participants

The sample was conveniently selected and consisted of 15 participants, aged between 25 and 33 years, who varied in their experience with generative AI tools. Among the participants, six were male and night were female. Their educational backgrounds ranged from bachelor's degrees to PhD's. Additionally, all participants demonstrated English language proficiency above level C1 and used English in their daily lives.

A recruiting Google form was employed to gather essential participant information, ascertain their viewing of the tutorial video associated with the tool to ensure an adequate understanding, and obtain their consent for recording purposes.

3.2 Execution of the experiment

The experiment comprises four primary phases: the Pre-Task tutorial, Task Accomplishment, Post-Task Interviews, and Crowd-sourced evaluations.

3.2.1 Pre-Task tutorial

Before commencing the experimental tasks, each participant will undergo a comprehensive pre-task tutorial. This tutorial aims to familiarize participants with the tool, clarify the assigned tasks, and ensure they possess adequate knowledge of the subject, specifically about bikes. To facilitate participants' understanding and usage of the tool's features listed in the subsequent post-experiment short questionnaire, all features will be demonstrated to them within the tool itself. Additionally, participants will be provided with a comprehensive wiki page containing information on bike parts and types. This additional resource will further enhance participants' knowledge and comprehension of the subject matter. Furthermore, a preliminary task involving the creation of a vase for tulips will be assigned, enabling participants to engage with the tool while seeking assistance if needed actively.

3.2.2 Task Accomplishment

During the experimental task, participants will be given the "bike product designer" persona for the company "22-Century Bike". They were also given specific instructions to create one or more new bike designs. The designs should be feasible for manufacturing, avoiding unconventional elements such as triangle wheels while emphasizing novelty/uniqueness and being aesthetically pleasing. Participants will have the opportunity to submit multiple designs for evaluation.

Participants will be encouraged to use the think-aloud protocol to verbalize their thoughts and decision-making processes during the task. Using the Leonardo.AI tool, they will have 30 minutes to create bike designs that fulfill the criteria of feasibility, novelty/ uniqueness, and aesthetically pleasing.

The given task instruction is as follows:

Task:

Using Leonardo.AI to create feasible, unique/novel, aesthetically pleasing bike design(s)

Instructions:
You work for the company 22-century Bike as a bike product designer. Your job is to make one or more new bike designs. Your new bike design(s) should be feasible to manufacture (no triangle wheels!) and as unique/novel and aesthetically pleasing as possible. You can submit multiple designs.

Remember to think aloud! The time will be about 30 mins.

3.2.3 Post-Task Interviews

The short questionnaire administered in this study aimed to gather valuable insights from participants regarding their experience using the Leonardo AI, which resembles the generative AI tool in this study. and its various features before the open-ended questions were asked. The questionnaire consisted of three sections, each utilizing Likert scale ratings.

The first section assessed the 'ease of use' and 'importance' of different features within the tool, utilizing a 5-point Likert scale ranging from "Not easy at all" to "Very easy." Participants were prompted to rate the ease of use for features such as text prompt, negative prompt, prompt generation, image input, guidance scale, upscale image, remove background, use generated image as input, edit canvas, and adjust image dimensions.

The second section evaluated the ease of creating feasible designs, novel/unique designs, and aesthetically pleasing designs using the tool. Participants provided ratings on a 5-point Likert scale ranging from "Not easy at all" to "Very easy."

The third section allowed participants to express their subjective opinions and perceptions regarding the quality of their bike design(s), encompassing feasibility, uniqueness, and aesthetics considerations.

After finishing the questionnaire, three open-ended questions were asked to gather participants' feedback on their experience with the tool. Participants are invited to provide a brief overview of their experience using the tool, highlighting both

positive aspects and areas for improvement. Furthermore, participants are asked to reflect on any challenges or difficulties they encountered while utilizing the tool and provide a description of these issues. Lastly, participants are given an opportunity to suggest additional features or functionalities that they believe would enhance the tool's performance and user experience.

The questions are:
1. Can you briefly describe your experience using Leonardo.AI? What did you like about it? What did you dislike about it?
2. Did you encounter any difficulties using the tool? If so, can you describe what they were?
3. Are there any additional features or functionalities you would like to see added to the tool to improve?"

3.2.4 Crowd-sourced evaluations

After completing the previous stages, the final set of generated images, consisting of 18 images, was collected based on the participants' decisions. Crowd-sourced evaluations were conducted to assess the quality of these images in terms of feasibility, novelty, and aesthetics. This approach involved gathering feedback from a diverse pool of participants, allowing for a broader range of perspectives and opinions to be considered during the evaluation process.

The crowd-sourced evaluations were conducted using Google Forms, and participants were recruited from the Prolific platform. The evaluation questionnaire included ratings for each bike image based on its feasibility, novelty, and aesthetics, using a 5-item Likert scale ranging from 'Strongly Disagree' to 'Strongly Agree'.

A total of 10 participants took part in the evaluation process, and on average, they spent approximately 6 minutes and 30 seconds rating the images.

3.3 Data Analysis

Since all the experiments were recorded, the first step involved transcribing the videos into a timeline format. This process involved converting the content of the videos into a chronological representation, allowing for easier analysis and reference.

The information regarding the timestamps of the prompts used, whether they were in global or local editing, and the specific criteria (feasibility, novelty, aesthetics) targeted by each prompt is compiled and listed in a file.

The Local/Global labeling is determined based on whether users utilize the prompt on the image generation page, which generates images in batches, or the AI canvas, which enables the modification of specific parts of an image. These interfaces are available within Leonardo.AI and can be seamlessly switched between based on the user's needs.

For the labeling of specific criteria such as feasibility, novelty, and aesthetics of prompts, we established the following definitions:

- Feasibility: explicit words related to manufacturing, addition, or modifications of bike parts related to the usability (whether it successfully functions/can be used) and/or manufacturability
- Novelty: explicit words related to uniqueness/novelty, prompt for any bike parts that are not in traditional bikes
- Aesthetics: any words related to the look/visual feel/dimensions of the bike

To label the prompts, we employ a collaborative approach involving myself and two PhD students who provide guidance on this project. We mark the prompts separately according to the defined criteria and ultimately determine the final labeling based on the majority agreement. Multi-criteria labeling is allowed. For example, if a prompt is relevant to both feasibility and novelty, it will be labeled with both criteria. When negative prompts are involved, we consider the main prompt along with the negative prompt. For instance, if the main prompt is primarily about feasibility and the negative prompt relates to aesthetics, we label that specific prompt iteration as both feasibility and aesthetics.

Based on the contents of this file, a spreadsheet is constructed, encompassing an extensive array of attributes that include quantitative and qualitative data. These attributes contain the basic participant information collected during the recruiting process, the generated images, prompts, labeling about the prompts (including whether they were in global or local editing mode and the criteria they focused on), sequence of prompt activity, the time spent on global and local editing, the time dedicated to each criterion, data obtained from post-task questionnaires, and the results of crowd-sourced ratings for each bike image.

The details about data visualization, correlation calculation, and text analysis will later be illustrated in the following section.

3.3.1 The Quantitative Data

The quantitative data encompasses various aspects, including the data from task execution, data obtained from post-task questionnaires, and crowd-sourced rating results. To facilitate subsequent analysis, all this data is encoded into numerical formats.

Furthermore, the analysis is enriched by additional attributes that track transformed or calculated information. These attributes include the percentage of time spent (by turning the time duration into percentage), the number of prompts used, the word count of the prompts utilized, and the average word count of prompts used in both global and local editing contexts, as well as for the three criteria. Other possible interesting attributes are also added; for example, "focus one or more than one criterion in the global prompt" is also a possible interesting attribute; the images/users are also given a label mono-criteria (0) or multi-criteria (1). This is given by finding the average number of criteria focused on global editing and splitting it into two groups using the mean as the threshold.

By incorporating these additional attributes, the data becomes more expansive, and we expect it to give us a deeper understanding of the experimental data, enabling more insights and meaningful findings.

Data Visualization

To gain insights into user behavior and identify potential patterns that contribute to generating better results, a variety of charts were employed first to visualize the data.

The first type of chart utilized is a time-line chart like fig 3.1a, where participants' approaches are represented on a single chart and ordered based on the crowd rating score of three criteria. This visualization provides an overview of the distribution of user behavior across the criteria and enables the identification of common patterns or trends.

The second type of chart focuses on the comparative analysis of attributes between the top-rated image and the bottom-rated image, as fig 3.1b shown. It comprises

two distinct groups of bars, each representing various attributes, including the percentage of time allocated to global and local aspects, feasibility, novelty, and aesthetics, as well as the number of prompts or the average word counts employed in each respective condition. These attributes are presented separately in the chart. This chart enables us to visualize the distinct approaches adopted by users for the top-rated image compared to the bottom-rated ones.

Finally, heat maps like fig 3.1c were utilized as a visual tool to depict the correlation between multiple numeric attributes. The correlation value is computed using Pearson's correlation coefficient. This graphical representation identifies positive or negative correlations between different attributes, thereby offering valuable insights into the interrelationships within the data. This analysis enables us to validate the observations made during the preceding stage and subsequently proceed with a statistical examination if significant correlations are detected.

Statistical Check

Upon identifying potential relationships between attributes, two primary methodologies were employed.

Firstly, to assess significant differences in values between two groups, such as the top-rated group versus the bottom group or the group favoring a focus on a single criterion versus multiple criteria, the t-test was employed.

The t-test is a powerful statistical tool that allows for the examination of whether there exist statistically significant differences in attributes between distinct groups. Our study enabled comparisons between different groups, including the top-rated and bottom-rated groups, or groups categorized by their preference for focusing on a single criterion versus multiple criteria. The formula for the t-test is as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two groups, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes of the two groups, respectively. By calculating the t-value and comparing it to the critical value at a chosen level of significance (usually set at 0.05), we could determine if the differences observed between these groups were likely to be due to chance or if they were truly meaningful and not a result of random variation.

Secondly, to explore the presence and strength of correlations between attributes, Pearson's correlation coefficient was utilized. This analysis helped determine whether any attributes positively or negatively influenced the final rating.

Pearson's correlation coefficient (denoted by r) measures the linear relationship between two continuous variables. It is calculated as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}},$$

where x_i and y_i are individual data points, $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively. The correlation coefficient r ranges from -1 to +1, where -1 indicates a perfect negative linear relationship, +1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship. By calculating Pearson's correlation coefficient between different attributes, we could assess whether they were positively correlated, meaning they tended to increase or decrease together or negatively correlated, indicating that one attribute tended to increase while the other decreased.

By employing these two methodologies, we were able to gain a comprehensive understanding of the relationships between different attributes in our study and determine whether they had a significant impact on the final rating.

3.3.2 The Qualitative Data

In terms of qualitative data, two main parts are considered. The first part comprises the prompts used by the participants during the image generation process, and the second part consists of the participant's answers to the three questions in the post-experiment interview.

Analyzing this data aims to identify commonly used terms frequently appearing across participants' prompts; for example, do participants tend to think of certain objects or ideas when entering prompts when they want to improve certain criteria? And are there any terms that help to generate better results?

Prompts analysis

To achieve the research objective, we utilized the term frequency-inverse document frequency (TF-IDF) measure, separately computed for two distinct files. Each file contains three columns corresponding to prompts categorized under feasibility,

novelty, and aesthetics, respectively. One file compiles prompts from the top-rated results based on specific criteria, such as the feasibility column containing prompts that received a top-rated feasibility rating. Conversely, the other file collects prompts from the bottom-rated results.

The TF-TDF is a numerical statistic commonly used to evaluate the importance of a term within a collection of documents. It involves two components: the term frequency (TF), which measures how frequently a term appears in a document, and the inverse document frequency (IDF), which accounts for the rarity of the term across the entire document collection. The TF-IDF value is calculated as follows:

$$\text{TF-IDF} = \text{TF} \times \text{IDF},$$

where

$$\text{TF} = \frac{\text{Number of occurrences of the term in the document}}{\text{Total number of terms in the document}},$$
$$\text{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term}}\right).$$

The TF-IDF value increases when the term appears frequently in the document and decreases when it appears frequently in other documents. This method is crucial in determining a term's significance when identifying different documents.

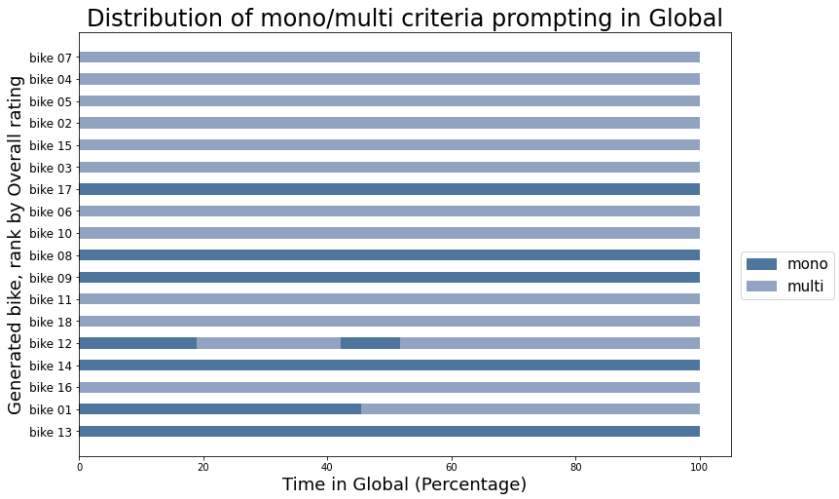
By calculating the TF-IDF values and generating a list of associated words, clear distinctions become observable. These findings will be presented and discussed in a subsequent chapter, providing a comprehensive understanding of the research outcomes.

Post-experiment interview answers analysis

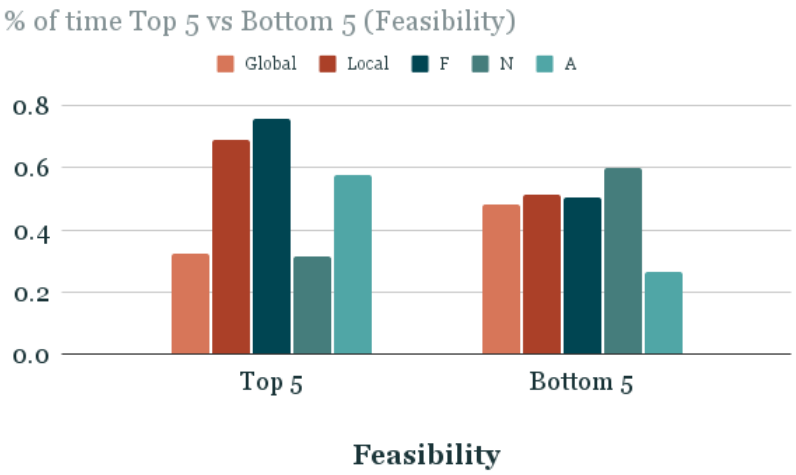
Furthermore, our analysis extends beyond quantitative ratings of ease of use and feature importance. By examining participants' responses to the three post-experiment interview questions, we aim to gain valuable insights into their perspectives on the user experience of using this prompt-to-image generative AI tool.

This qualitative approach strives to uncover participants' opinions and suggestions regarding the overall improvement of these tools, focusing on their experiential aspects. Through these interviews, we seek to understand how users perceive the functionality and how they emotionally connect with and experience the tool. This deeper understanding will be essential in refining and enhancing the user

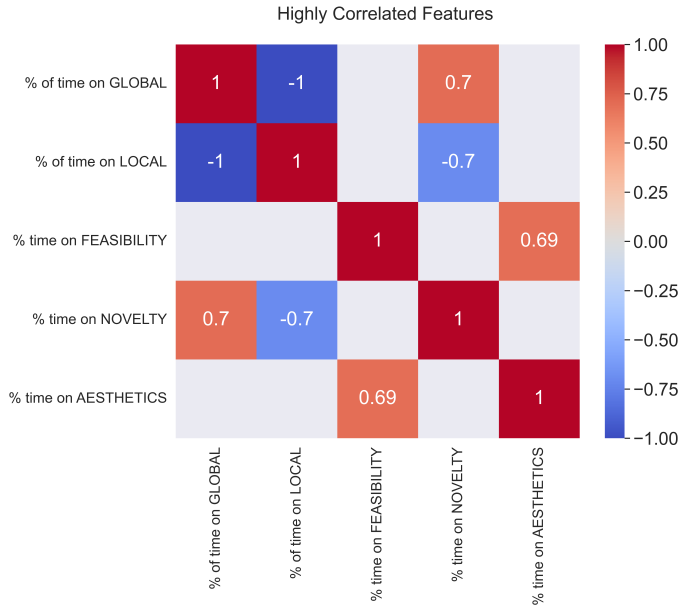
experience, ensuring that future AI tool iterations can be more effective for user needs and preferences. The qualitative analysis will complement the quantitative data, comprehensively evaluating the tool's performance and usability.



(a) Example of the timeline chart of each user



(b) Example of the comparison chart between groups



(c) Example of heat map of correlation

Fig. 3.1: The three main types of virtualization of collected data

4 Results

This chapter thoroughly compiles the key findings from the thorough analysis conducted in the previous section. It encompasses the outcomes obtained through data visualization, correlation, and text analysis.

4.1 Exploring User Approaches in Designing Products with Prompt-to-Image Generative AI Tools

The first question we want to answer through the result is, "How are users currently using prompt-to-image Generative AI tools to design products that meet their requirements?" In order to find out the answer, we have identified several commonly employed approaches among the majority of users.

4.1.1 Users start with global prompts and move to local refinements, spend most of their time on local

One notable aspect of user behavior with respect to prompt-to-image Generative AI tools is how users choose to allocate their time between global and local editing. As a reminder, global editing means that users try out prompts that change the entirety of the AI-generated image, whereas local prompt editing means that they change only a part of this image. The visualization of time distribution is shown in fig 4.1.

Through our observations, it became evident that most users (14 out of 15) engage in both global and local editing, while only one user exclusively focuses on global editing. It is expected that all users initially start with global editing, experimenting with several tries of prompts (on average, 4.36 tries) before selecting an image for refinement.

On average, users allocate approximately 1/4 of their time to global editing, concentrating on broader modifications, while the remaining 3/4 is dedicated to local editing, refining specific details to meet their requirements. Furthermore, we noticed

a tendency among users who enter the local refinement mode to revert back to the global editing mode rarely. Only 3 out of 15 users switched back to global editing after entering the local mode. This behavior aligns with the concept of "design fixation" [JS91] [Lin+10], where designers overly focus on their initial ideas without considering changes, often resulting in compromised quality. However, further statistical analysis would be necessary to examine this observation rigorously.

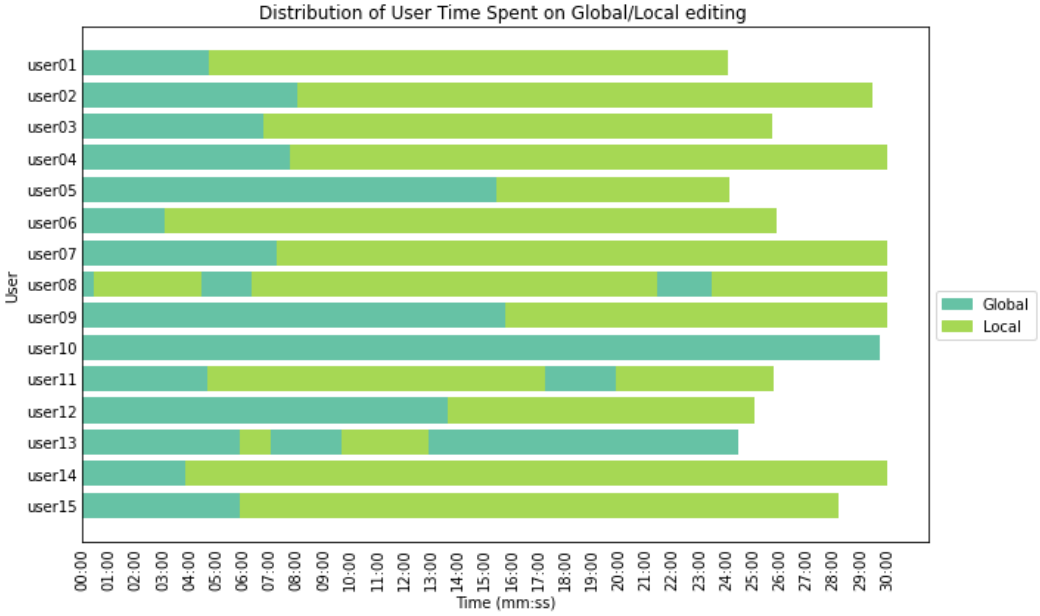


Fig. 4.1: The distribution of User Time Spent on Global/Local editing, we can see that most of the users spend more time on the local than the global phase, and not many users jump back and force from local to global.

4.1.2 On global, users focus on novelty, while in the local mode, they focus on feasibility

While checking the correlation heat map fig 4.2, we noticed that the percentage of time spent on global has a high correlation with that spent on improving the novelty, and the fig 4.3 shows that prompts used on global also has a high correlation with that used to improve the novelty.

A Pearson correlation analysis was performed to investigate the relationship between the percentage of time spent in the global phase and the percentage of time spent on novelty. The correlation coefficient $r = 0.70,\ p < .05$, indicates a moderate positive correlation and statistical significance. The same test of the relationship between the number of prompts used in the global phase and those employed to improve novelty

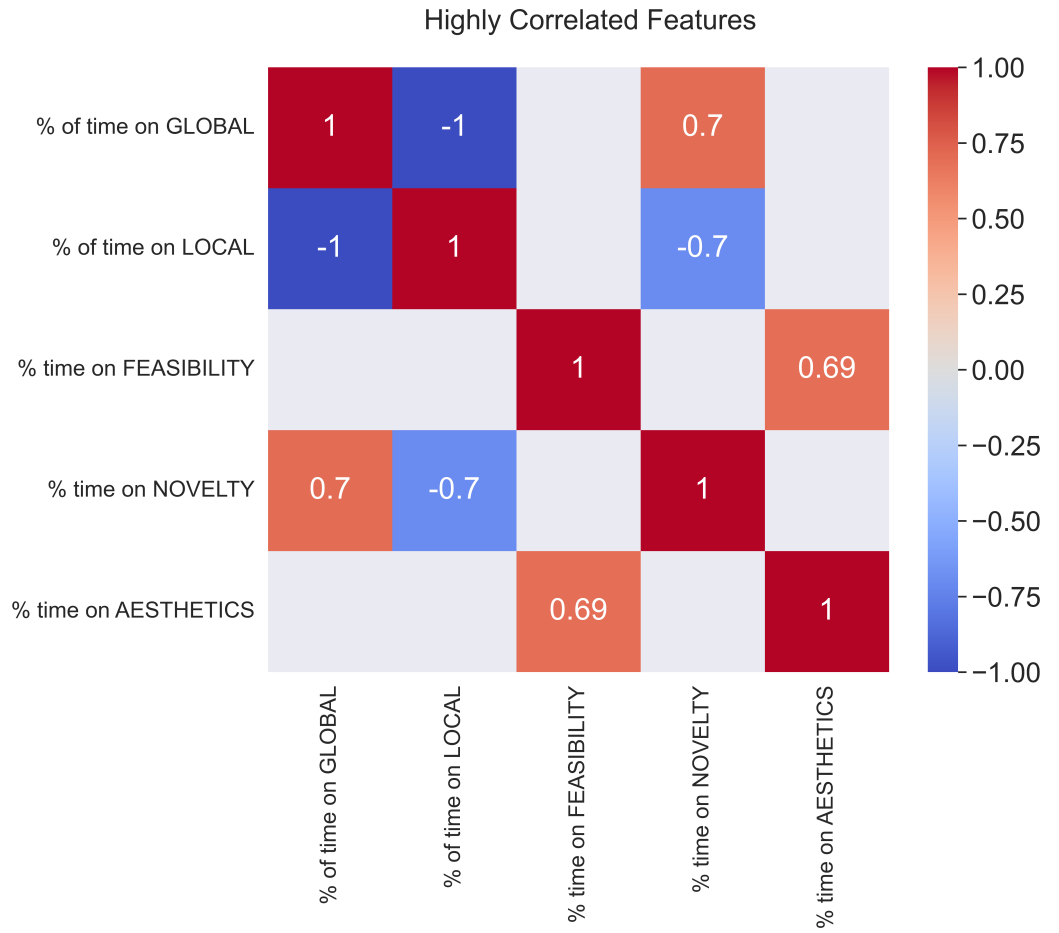


Fig. 4.2: The heat map of the percentage of time spent on global, local, Feasibility, novelty, and aesthetics. The Correlation value of the percentage of time spent on novelty and the percentage of time spent on the global phase is high ($r = 0.7$)

revealed a correlation coefficient $r = 0.72,\ p < .05$, indicating a strong positive correlation between these variables and statistical significance.

These results demonstrate a significant positive correlation between both the percentage of time spent in the global phase and the percentage of time spent on novelty and the number of prompts used in the global phase and those employed to improve novelty. In other words, as individuals allocate more time or try more prompts in the global phase, they also tend to allocate a higher percentage of their time and number of prompts to novelty.

Although the correlation heat map did not indicate a significant relationship between the percentage of time spent on local editing and any of the criteria, we can see an obvious difference in the number of prompts that focused on feasibility used in the global phase and local phase in fig 4.4. And there was a noteworthy correlation

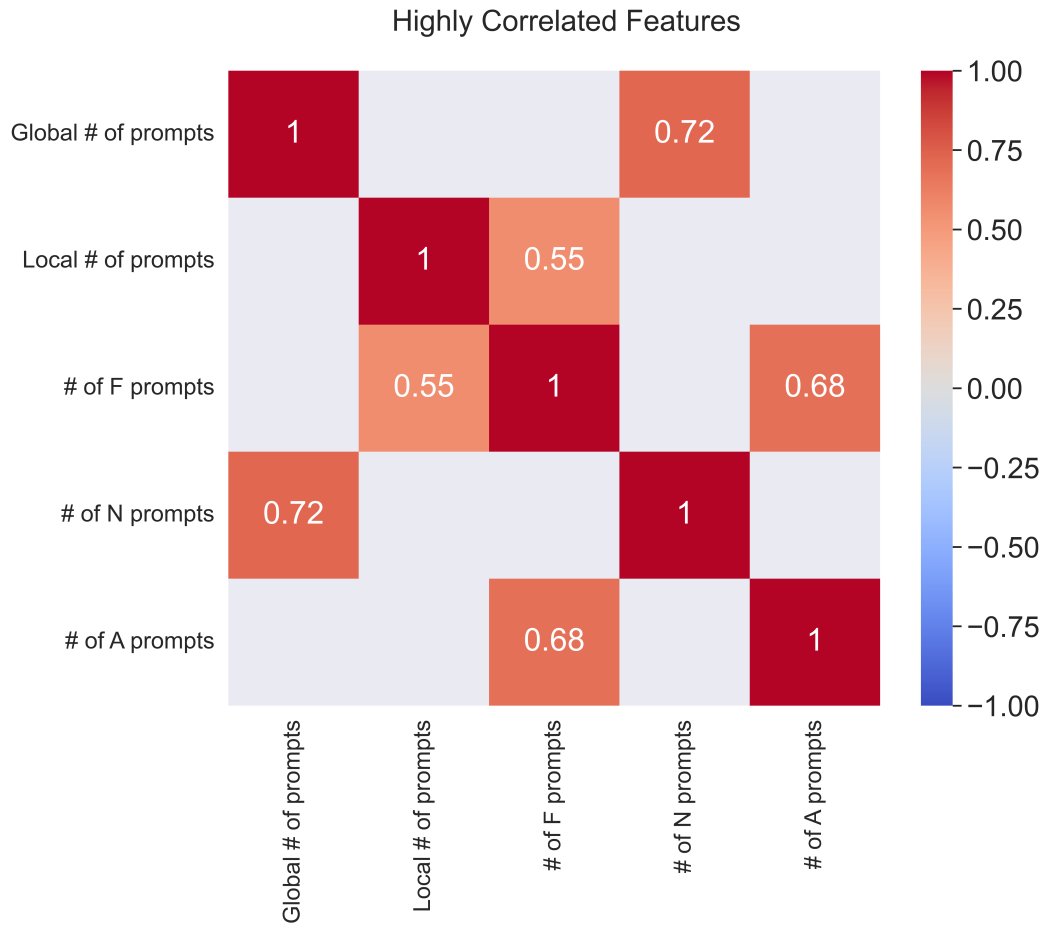


Fig. 4.3: The heat map of the number of prompts utilized on global, local, feasibility, novelty, and aesthetics. A correlation was observed between the number of prompts used for local refinement and those employed to improve feasibility. ($r = 0.55$)

observed between the number of prompts used for local refinement and those employed to improve feasibility. The correlation analysis revealed a coefficient $r = 0.56,\ p < .05$, suggesting a moderate positive correlation between these variables and indicating statistical significance.

The possible reason might be that the global phase involves a broad evaluation of multiple criteria, fostering the exploration of novel ideas and perspectives. In contrast, the local phase focuses on refining the design based on practical considerations. The cognitive processes involved in evaluating novelty and feasibility differ, leading to a natural emphasis on different aspects in each phase.

On the other hand, novelty is a more complex criterion than feasibility. It is difficult to define what makes an image novel, and it can be difficult to provide instructions to the AI model that will result in a novel image. As a result, users may need to

spend more time and effort in the global editing phase to achieve their desired level of novelty.

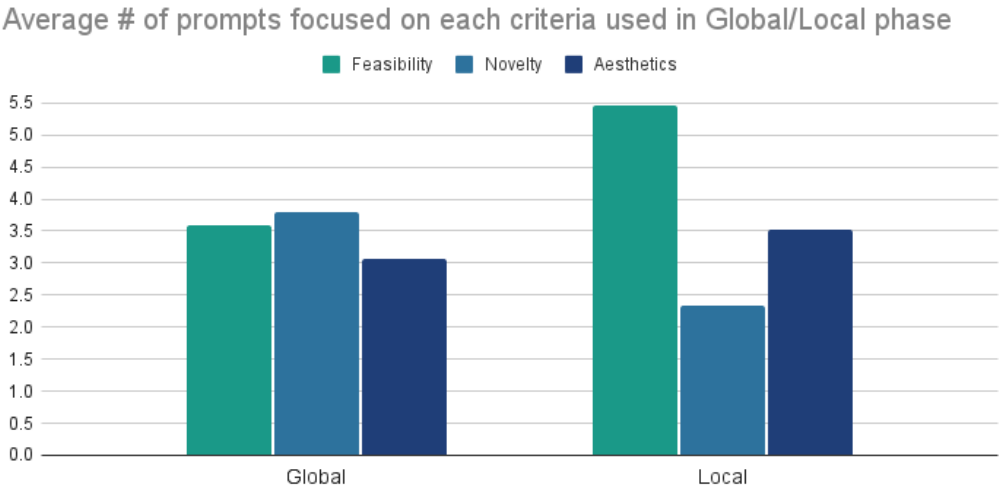


Fig. 4.4: The average number of prompts focused on each criterion used in the Global/Local phase, we can see an obvious difference in the number of prompts that focused on feasibility used in the global phase and local phase

4.1.3 User write lengthier prompts in Global Editing Phases and shorter prompts in Local editing phase

Our analysis reveals a distinct pattern in prompt length across the editing phases shown in fig 4.5. Users tend to employ longer prompts during global editing, possibly providing more comprehensive instructions to the AI model. However, during local editing, users opt for shorter prompts, indicating a shift towards focusing on specific details and refining the generated image based on their requirements.

The reason is not just because that users tend to focus more criteria on the global phase. If we divide the average word count by the criteria focused, on average, the prompt in the global phase is still more extended. Some possible reasons include:

- Global editing is a more complex task. In the global editing phase, users typically try to give the AI model a general overview of what they want the image to look like. This can involve a lot of different criteria, such as the overall composition, the mood, the style, and the objects that should be included. As a result, users may need to provide more detailed instructions to get the desired results.

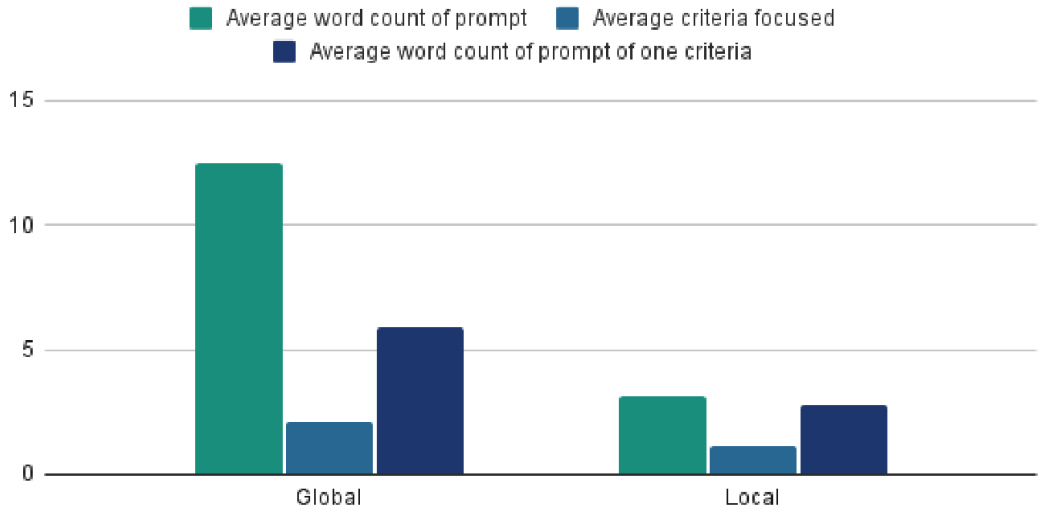


Fig. 4.5: The average word count of prompts and average criteria focused on, of the prompts in the Global/Local phase, we can see that the average words used in global prompts are much higher than in local

- Local editing is a more focused task. In the local editing phase, users are typically trying to refine specific details of the image. This might involve things like changing an object's color, a person's position, or the size of a text element. As a result, users may be able to get by with shorter prompts that are more specific to the changes they want to make.
- Users may be more aware of the limitations of the AI model in the global editing phase. In the global editing phase, users typically try to create a new image from scratch. As a result, they may be more aware of the limitations of the AI model and the need to provide more detailed instructions. In the local editing phase, users typically try to refine an existing image. As a result, they may be more confident in the ability of the AI model to make small changes and may be less likely to provide detailed instructions.

4.1.4 Shared and Synergistic prompting between Feasibility and Aesthetics

Back in fig 4.2 and 4.3, we notice the high correlation between the number of prompts used and the percentage of time spent on feasibility and aesthetics as well. The correlation analysis revealed a coefficient $r = 0.69$ and 0.68, respectively, suggesting a positive correlation between these variables. Furthermore, the associated $p < .05$, indicating statistical significance.

In the fig 4.6 also shows that the number that focused on both feasibility and aesthetics is the most, much high than the combination of feasibility + novelty and novelty + aesthetics.

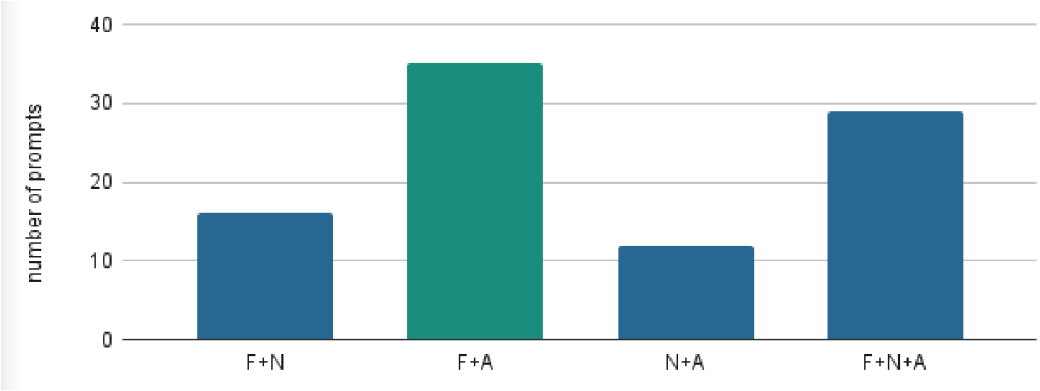


Fig. 4.6: The number of prompts that focused on multiple criteria, we can notice that the F+A labeled prompts are more than other combinations

After the prompt analysis, it appears that users instinctively employ prompts that improve both feasibility and aesthetics. This natural inclination showcases the users' intrinsic understanding that design should not only be visually appealing but also practically implementable. Users intuitively use prompts that take care of both aspects, ensuring that the generated images meet their functional requirements while exuding an aesthetic appeal. This result can also be confirmed by prior studies, which refer to the "aesthetic-usability" or "aesthetic-utility" effect [KK95] [SS10], which underscores the phenomenon where people perceive aesthetically pleasing designs as more usable and effective, even when functionality remains unchanged.

4.1.5 Common topics in people's prompts

Participants often focused on specific bike parts that would enhance the feasibility and functionality of the design, like word cloud fig 4.7a shows. Some of the recurring prompts related to feasible components include the tube, pedal, saddle, and wheels. For example, one participant wrote the following prompt: "Regular top tube down tube and pedal that makes it look like a legit bike."

For novelty, as word cloud fig 4.7b shows, they showed an interest in incorporating future technologies and innovative features into their designs. Notable examples of novel elements include electricity-powered mechanisms, battery systems, covers, LED lights, wings, and the exploration of new and sustainable materials like wood.

For instance, one participant wrote the following prompt: "Electrical bike with a large comfortable seat and streaming body."

In aesthetics prompts shown in word cloud fig 4.7c, participants expressed a keen interest in aspects like color, shape, and pattern on the bike. For example, one participant wrote the following prompt: "Please create a rainbow color bike design."

(a) The word cloud of feasibility prompts **(b)** The word cloud of novelty prompts **(c)** The word cloud of aesthetics prompts

Fig. 4.7: Feasibility, novelty, and aesthetics word clouds

4.2 Strategy of success for the Prompt-to-Image Generative AI Tools

Several key observations have emerged based on analyzing the crowd rating results and the various approaches users take. These observations can serve as recommended strategies for effectively leveraging generative AI tools such as Leonardo.AI. These strategies are as follows:

4.2.1 Allowing the AI to think for you? Also remember to direct it toward your specific creative design

During the experiment, we observed that some users prefer to let the AI tool take the lead in the design process, while another group prefers to use directive prompts.

The first group of users provided high-level prompts such as "a special bike" or "an aesthetically pleasing bike." By employing such open-ended prompts, these users encouraged the AI to exercise its creativity, allowing it to generate unique and innovative bike designs that aligned with their general preferences. This collaborative

and exploratory interaction enabled the AI tool to become a co-creator, shaping the final output in partnership with the users.

On the other hand, the second group of users adopted a directive style, providing specific and precise prompts like "a bike with square wheels." or "Electrical bike with a large comfortable seat and streaming body". These users conveyed their concrete design requirements with explicit instructions, leaving relatively little room for ambiguity. By taking this directive approach, they sought to guide the AI tool in producing bike designs that precisely matched their envisioned concepts. In this case, the users' role resembled that of an informed supervisor, overseeing the design process and ensuring that the outcome aligned precisely with their preconceived ideas. Notably, during our study, we observed a balanced distribution of users between these two styles. Approximately half of the participants preferred high-level, exploratory prompts, while the other half opted for directive, specific prompts.

To compare if one group has better results, a t-test is conducted. The bike designs are divided into two groups:

Group 1: Bike designs that users use high-level, exploratory prompts
Group 2: Bike designs that users use directive, specific prompts

A t-test was conducted to examine whether there is a significant difference in the crowd rating of the feasibility of the bike designs between users who use high-level prompts ($M = 3.0$) versus users who use the directing one ($M = 3.45$). Results indicated that there is a statistically significant difference between these two user groups, with $t(16) = 2.22, p < .05$. However, no significant differences were found in the crowd ratings for the novelty and aesthetics of the bike designs between the two groups.

This discovery indicates the interplay between user involvement and the AI's creative process in design tasks. While both approaches, high-level prompts, and directive prompts, can yield aesthetically pleasing and novel bike designs, the user group that employed directive prompts achieved higher feasibility ratings. As a result, this finding suggests that when users provide clear and detailed instructions, they effectively hold the AI's creativity within practical boundaries, leading to designs that are imaginative and highly viable for real-world applications. By offering specific guidance, users help the AI to achieve its potential and incorporate essential elements, ensuring that the generated designs stand in the right balance between creativity and functionality. This user-led approach empowers the AI to produce innovative solutions with a higher probability of successful implementation in the physical world.

4.2.2 Multi- vs. Mono-Criteria Prompts: Broad Start is more Beneficial

During the experiment, we noticed that some users employed prompts that covered multiple criteria, such as "square tire bike with bottle cage and red chain ring," while others concentrated on a single criterion (mono-criteria), such as "a bike with special design." To dig deeper into the potential impact of this behavior on the outcome, we analyzed the timeline charts of the Multi- vs. Mono-Criteria prompts used on global editing of the bikes shown in Figure 4.8 . The bike in the charts ranks according to average crowd ratings, allowing us to identify any noticeable patterns.

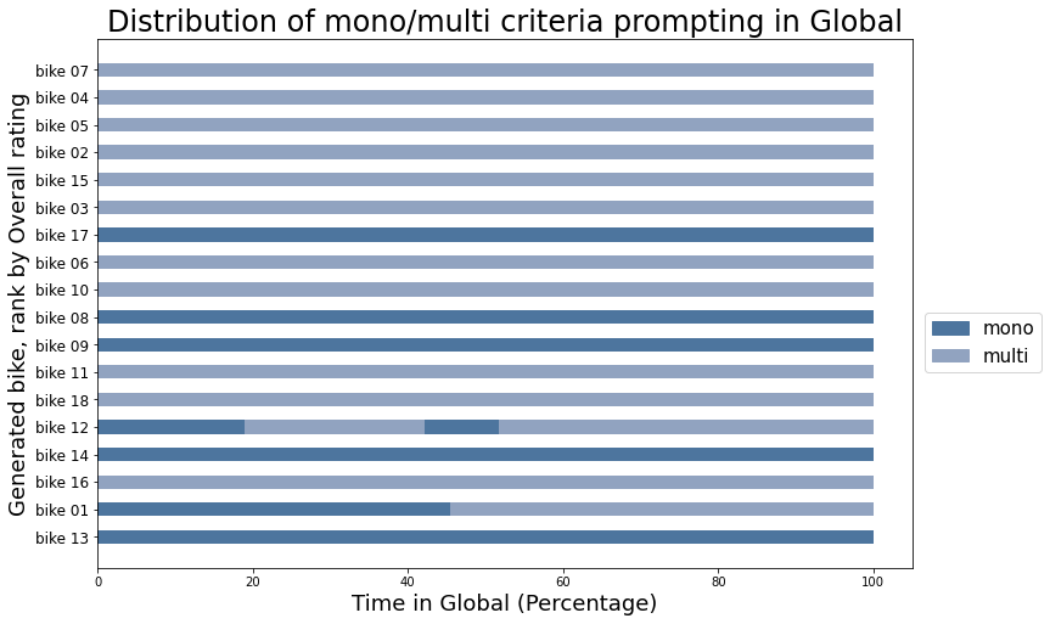


Fig. 4.8: The distribution of user time spent on each criterion, ranked by the crowd's feasibility rating from top to bottom. We can see that the users that start with mono-criterion are often listed in the bottom

When comparing the disparity between the top and bottom-rated users, a notable trend is observed in the feasibility ranking shown in Figure 4.8. Most top-rated users begin with multi-criteria prompts, while the bottom-rated users predominantly start with mono-criteria prompts.

To validate this observation, a t-test is conducted. The users are divided into two groups:

Group 1: Users who use the majority of mono-criteria prompts in the global phase
Group 2: Users who use the majority of multi-criteria prompts in the global phase

A t-test was conducted to examine whether there is a significant difference in the average crowd rating quality of the bike designs between users who use the mono-criteria approach ($M = 2.95$) versus users who use the multi-criteria one ($M = 3.37$). Results indicated that there is a statistically significant difference between these two user groups, with $t(16) = -2.13, p < .05$

This result indicates that users who initiate their evaluations in global editing using more multi-criteria prompts tend to achieve better results. The observed trend suggests that starting with multi-criteria prompts enables users to consider and communicate a broader range of design attributes, leading to more comprehensive and well-rounded bike designs.

The preference for multi-criteria prompts aligns with the principles of user-centered design, where a diverse range of perspectives and factors is considered to create a well-balanced and user-preferred end product. By incorporating various design attributes into the prompt, users effectively guide the generative AI tool in producing designs that are not only aesthetically pleasing but also functionally and practically sound.

4.2.3 To Boost novelty, try more novel-related prompts, but more feasibility-related prompts do not guarantee better feasibility

When examining the rankings, we noticed an interesting trend - users who received high novelty ratings tended to use more prompts specifically targeting novelty. We conducted a Pearson correlation analysis to explore this relationship further and ascertain its statistical significance, and the result can be seen in fig 4.9.

The Pearson correlation coefficient between the number of novelty-focused prompts used by users and their novelty ratings was found to be $r = 0.58, p < .05$. This indicates a moderate positive correlation between the two variables. The statistically significant p-value < 0.05 suggests that this correlation is unlikely to have occurred by chance, providing strong evidence of the relationship.

In contrast, our analysis of the correlation between the number of feasibility-related prompts and the feasibility rating revealed no significant associations. The correlation coefficient was close to zero, indicating a minimal to no linear relationship between the number of feasibility-focused prompts.

Based on the analysis conducted, it appears that using more novelty-focused prompts can significantly boost novelty ratings. Therefore, if you want to enhance novelty

in your prompts, it would be beneficial to use more prompts specifically targeting novelty-related aspects.

However, the analysis did not find a significant correlation between the number of feasibility-related prompts and feasibility ratings. This suggests that simply using more feasibility-related prompts may not guarantee better feasibility ratings. To improve feasibility, one of the possible ways is discussed in the following paragraph.

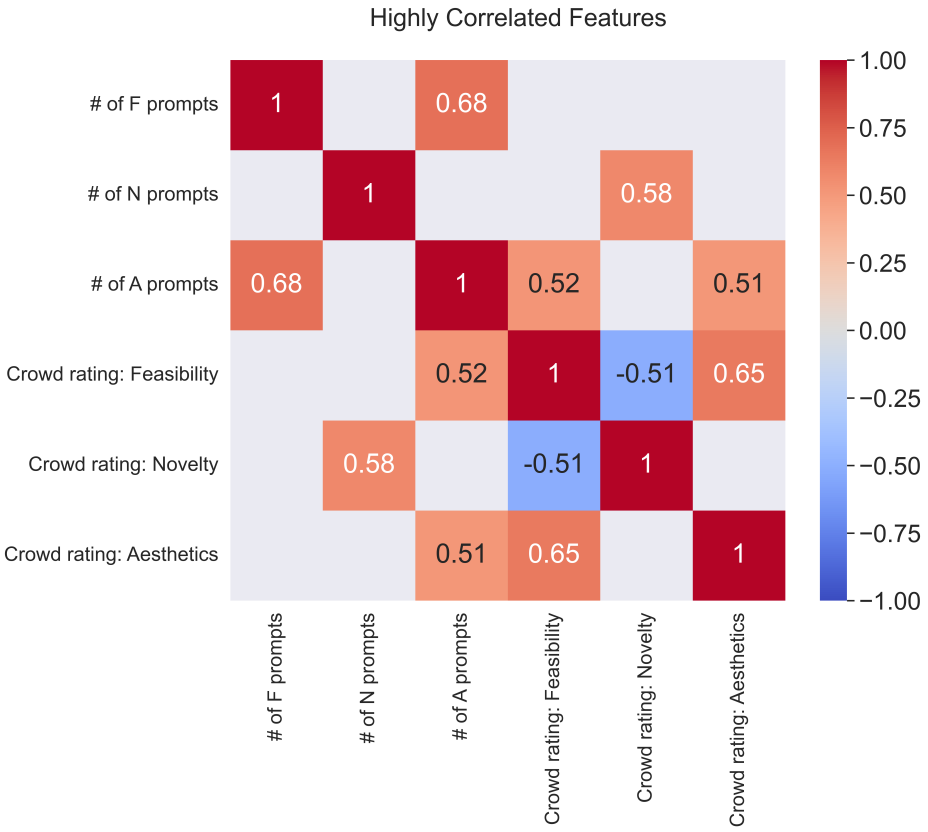


Fig. 4.9: The number of prompts that focused on each criterion. The Pearson correlation coefficient between the number of novelty-focused prompts used by users and their novelty ratings was found to be $r = 0.58, p < .05$

4.2.4 The Power of Aesthetic Prompts to Enhance Feasibility (as well as Aesthetics)

In fig 4.9, we can observe another intriguing finding: the number of aesthetics-related prompts positively correlates with the feasible crowd rating. The results of the Pearson correlation analysis revealed that using more prompts that focus on aesthetics not only improves the aesthetics rating (Correlation coefficient $r =$

$0.52, p < .05$) but also positively impacts the feasibility rating (Correlation coefficient: $r = 0.51, p < .05$). These correlations suggest that considering aesthetics during the design process can have a beneficial influence on both the visual appeal and practicality of the generated designs.

Continuing with the research findings in 4.1.4, the positive correlation between aesthetics-related prompts and both the aesthetics and feasibility ratings suggests that aesthetics play a significant role in enhancing the overall quality of designs. When incorporating aesthetically pleasing elements, not only are the visual preferences of users satisfied, but the practicality and functionality of the designs also improve. This synergistic relationship between aesthetics and feasibility is often referred to as the "aesthetic-usability" or "aesthetic-utility" effect.

Considering this "aesthetic-usability" or "aesthetic-utility" effect, it is recommended for users involved in design processes to incorporate aesthetically pleasing elements into their designs actively. By doing so, not only will the visual preferences of users be satisfied, but the practicality and functionality of the designs will also improve. This approach can lead to more appealing and user-satisfied designs that are likely to be well-received by the target audience.

However, if the design is intended for physical manufacturing, it becomes even more crucial to consider additional tools that incorporate 3D modeling or physics engines. These tools can significantly boost the feasibility of the design and provide a more direct and realistic representation of the final product.

4.2.5 The More Is Not Always the Merrier: Lengthy Prompts Aren't Always the Key to Success

Besides what users are focused on for each prompt, we are curious whether providing more details, probably longer prompts, always leads to better results. However, our analysis challenges this assumption. While longer prompts may seem comprehensive for humans, they do not consistently improve the outcomes.

To dig deeper, we conducted a correlation analysis (Pearson correlation) between the mean word count of prompts and the corresponding ratings. Surprisingly, our findings reveal no discernible relationship between the prompt length and the ratings of feasibility, novelty, and aesthetics. This holds true whether we consider the mean word count of the global prompt, local prompt, or overall prompt (the p-value of all combinations is not less than the commonly used threshold of 0.05).

These results indicate that multiple factors contribute to the quality of the generated output. While clarity in prompts remains crucial, brevity does not necessarily hinder the potential for excellent results. Building on the findings discussed in section 4.2.1, it appears that employing specific prompts can enhance idea generation more effectively than using high-level prompts alone.

Based on the analysis and findings presented, the user should focus on clarity and specificity when providing prompts rather than simply aiming for longer ones. While longer prompts may seem comprehensive to humans, they do not consistently lead to better feasibility, novelty, and aesthetic results.

Instead, users should consider employing specific prompts to enhance idea generation more effectively. Incorporating targeted and detailed text within the prompt can be more advantageous in achieving the desired outcomes. It is not about the length of the prompt but rather the quality and relevance of the information provided.

As a result, when using the AI system, try to be clear and precise in your prompts, for example: "electric bike with a low saddle ". Do not make your prompts overly lengthy, for example: "Please create a bike design that is feasible to manufacture and function as a bike for humans e.g., no triangle wheels and with handle, and as unique/novel and aesthetically as possible. " will not be a good way of prompting, at least in the current AI model.

4.2.6 Decoding the Keywords of Successful Prompting

As longer prompts do not guarantee to generate better results, to understand what truly contributes to success, we delve deeper into the disparity between the prompts used by the top five users and the bottom five users. We aim to discern if a shared set of keywords or common concepts is key to their achievements.

By calculating the TF-IDF (Term Frequency-Inverse Document Frequency) of the prompts from both groups of the top five users and the bottom five users based on the ranking of each criterion (feasibility/novelty and aesthetics), we uncovered the standout words with the highest TF-IDF values within each category:

- Feasibility:

  Common keywords in the prompts of top feasibility rating design:
  Feasible, Manufactured, Add, Human, Create, Metal, Tube, Chain, Tire: These words suggest practicality, real-world viability, and ease of manufacturing and implementation in bike design.

Common keywords in the prompts of bottom feasibility rating design:
Future, Propeller, Hammock: These terms may indicate concepts that are difficult to implement or may not align with the feasibility of a bike design.
Sofa, Anti, Gravity, Comfortable: Words that might not be applicable or relevant to bike design in terms of feasibility.

- Novelty:

  Common keywords in the prompts of top novelty rating design:
  Propeller, Dive, Fly: These words suggest elements of aircraft or flying, indicating a creative and unique approach to bike design.
  Innovative, Future, Transparent, Electrical, Glasses: Terms that imply modern technology integration, power, or material to be applied, making the bike design novel and cutting-edge.

  Common keywords in the prompts of bottom novelty rating design:
  Wing: While wings and flight-related terms may be novel, they are not commonly applicable or practical for regular bikes.
  Phone, Screen, Holder, Speaker, Touch: Words that may imply integrating smartphone or electronic components into the bike, but they may not be seen as necessary or functional.

- Aesthetics:

  Common keywords in the prompts of Top aesthetic rating design:
  Bottle, Tube, Pedal, Cage: These words likely describe sleek and visually appealing components.
  Racing, Athlete: Terms that evoke a sense of speed and performance in the design.
  White, Black, Blue, Color: Color-related words that may signify a harmonious and attractive color scheme.

  Common keywords in the prompts of bottom aesthetic rating design:
  Decoration, Modern: These words suggest irrelevant or potentially conflicting design elements for a bike.
  Forest, Mountain, Lake: Nature-related terms that may not relate well to bike aesthetics. Crazy, Simple: Subjective terms that do not provide a clear indication of what makes the design aesthetically good or bad.

From the above observations, we can conclude with some guidelines and suggestions for good prompting.

When writing a prompt that aims for better feasibility, it is crucial to be specific and use language related to practical bike design. Focus on materials, components, and manufacturing processes commonly used in the bike industry. By avoiding conflicting features and emphasizing functionality, the generated designs are more likely to be safe and efficient for real-world implementation.

For the improvement of novelty, the prompt should request innovative and unconventional design elements. Consider integrating futuristic technologies or unique applications to push the boundaries of traditional bike concepts. Open-ended prompts can inspire diverse and imaginative ideas, encouraging the model to think outside the box.

About aesthetics, using expressive language in the prompt to guide the model in creating visually appealing designs. Specify desired color schemes, iconic features, and overall design themes for a distinctive appearance. Considering the target audience's preferences – whether they prefer a sporty, elegant, or urban-oriented style – can help generate designs that resonate with users.

Overall, a good prompt for achieving good quality of feasibility, novelty, and aesthetics takes work. However, being specific and concise is crucial to avoid ambiguity and allow the model to focus on critical aspects of the design. Additionally, providing relevant context or constraints can guide the model in tailoring designs for specific contexts or user preferences. Using these guidelines and customizing prompts to suit users' needs and focus should improve the quality of generated bike designs.

4.2.7 Be aware of the Trade-Off: Relationship Between Feasibility and Novelty

As fig 4.10 shown, we observed a negative correlation between the feasibility and novelty ratings of the generated results (Correlation coefficient $r = -0.51, p < .05$). This correlation coefficient of -0.51 indicates a moderate negative relationship between feasibility and novelty. The corresponding p-value of $< .05$ further confirms that this correlation is statistically significant. In simple terms, when the feasibility rating of a prompt increases, there is generally a decrease in the novelty rating, and vice versa. This finding emphasizes the trade-off between feasibility and novelty inherent in the generated outputs.

Given this observation, it is important for users to recognize this trade-off when using the generative AI tool. Understanding that feasibility and novelty often move in opposite directions allows users to set realistic expectations and make informed

decisions. Furthermore, users should determine their priorities and goals before generating prompts. They should consider whether they prioritize highly feasible ideas or if novelty and innovation are more important to them. Users can align their prompt generation efforts by clarifying their primary focus.

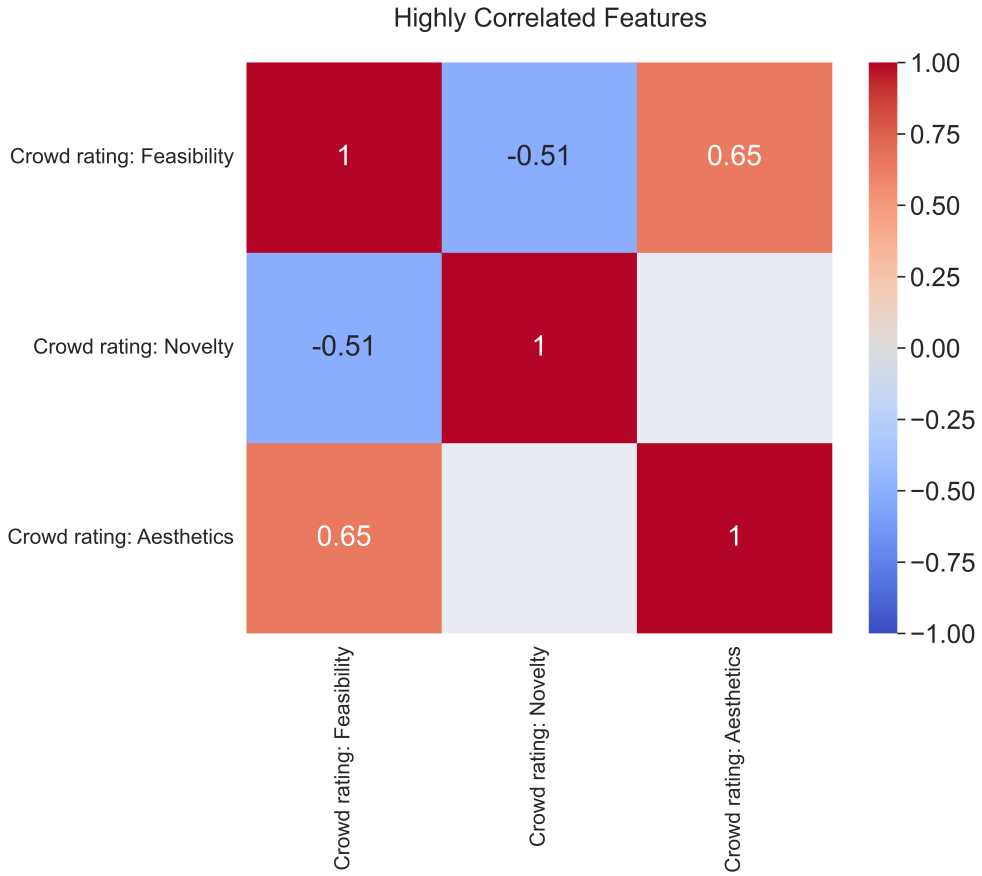


Fig. 4.10: The average number of prompts focused on each criterion used in the Global/Local phase. We observed a negative correlation between the feasibility and novelty ratings of the generated results (Correlation coefficient $r = -0.51, p < .05$)

5 Discussion, Limitations and Further Work

This chapter focuses on recommending features that can be added to the generative AI tool, which generates images based on user prompts. We will explore the usefulness of included features and discuss some points that require improvement or additional functionalities. Furthermore, we will address the limitations of the experiment design and provide insights for future enhancements.

5.1 Design Recommendations

Based on the post-task questionnaire, the feedback indicates that users generally found Leonardo.AI to be a functional and user-friendly tool. The tool received decent overall ratings for ease of use and the quality of the generated images, averaging around 3.5 out of 5. However, some specific functions were identified as less intuitive or less significant according to user feedback in the post-task questionnaires. These functions include the negative prompt (for specifying what shouldn't be included in the generated images) and image scaling (for adjusting the size of the image).

Most users primarily utilized the basic prompt-to-image function (global editing) along with the edit canvas feature (including an eraser and local prompting refinement) and modifying the guidance scale (adjusting the weight of the prompt in the model). Few users explored more advanced functions such as up-scaling the image or removing the background.

Regarding the basic functions, users reported similar issues, such as the tool not accurately following their prompts or the generated outcomes not meeting their desired results. User feedback included quotes like "The tool may not always accurately follow the given prompt, resulting in outputs that don't align with my intentions," "The tool fails to generate the specific things I want," and "Some commands cannot be executed correctly, and there are limitations in adding specific elements." More than one uses also reflect that the tool is slow or not quick enough. As a result, it appears that the design of this generative AI tool should prioritize the improvement

of the performance and user-friendliness of the basic functionality. Additionally, more tutorials and examples should be provided to encourage users to explore and make the most of the advanced functions.

By the collected post-task interviews, some additional features and functionalities users would like to see improved or added to the tool are:

Style customization such as filter

Style customization with diverse filters in the generated designs would significantly enhance the tool's appeal. Intuitive controls and a user-friendly interface make experimenting effortless, empowering users to add a personal touch to their creations even without prompting. This enhancement could transform the tool into a versatile artistic creation and exploration platform.

Smarter AI integration and improved language understanding

Enhancing the tool's comprehension capabilities, especially in understanding prompts and specific instructions, would bridge the communication gap between users and the tool. Improved language understanding would enable more accurate and precise design generation.

Partial adjustment options

Users want the flexibility to make partial adjustments to the generated designs, such as adjusting the size, direction, or rotation. This would provide more customization options and allow users to fine-tune the designs to their specific needs.

Image gallery integration

Integrating an image gallery into the tool would greatly enhance its capabilities, empowering users to search, access, and incorporate a diverse range of existing pictures and elements. Presently, users are limited to uploading entire images, but the proposed gallery would allow them to search for specific items or elements within the tool itself, streamlining their design process and significantly improving overall efficiency.

Three-dimensional design capabilities

Users desire the ability to work in a three-dimensional space, allowing them to create designs with depth and realism. This feature would enhance the tool's capabilities and enable users to explore three-dimensional design concepts.

These additional features and functionalities would improve the overall functionality of the tool and provide users with more control and customization options when generating designs.

5.2 Limitations

Understanding the limitations is crucial as they can have a potential impact on the interpretation and generalizability of the results. By being aware of these limitations, we can gain a better perspective on the scope and reliability of the findings. It also allows us to identify areas that require further investigation and possible improvements in future research.

Too few users included in the experiment

The small sample size that could result from the experiment's few participants could have an impact on how generalized the results are. The robustness and representativeness of the results would be improved by a larger and more varied user group.

Users are not experienced or familiar enough with the generative AI tool

Users may experience a learning curve due to their unfamiliarity or lack of experience with the specific generative AI tool used in the experiment. This might limit their ability to make full use of the tool's features and may affect the way they interact and provide feedback. Additionally, we found that the majority of users in our studies only used about half of the features.

Subjective Criteria

The assessment of design quality in terms of feasibility, novelty, and aesthetics is inherently subjective. While crowd-sourced evaluations help to gather diverse opinions, individual perceptions of design quality can vary. Future studies could explore additional objective criteria for assessing design quality or employ expert evaluations to complement user feedback.

Task Time Constraint

The 30-minute time constraint for completing the design task may have impacted the depth and complexity of the designs participants could create. Some participants may have felt rushed and were unable to explore all possible design options. Allowing for more extended task periods could yield more elaborate and innovative designs.

Image and terms of the bike might not be enough to feed into the back-end model training

During the execution of the experiment, it became evident that terms associated with bikes did not always help to elicit accurate results. This observation led us to consider that one of the potential factors contributing to this issue could be the limited set of images and terms utilized for training the back-end model. The model's performance and comprehension are likely hindered by the inclusion of excessively specialized or industry-specific terminology, which may not encompass a sufficiently broad and general range of bike-related concepts.

We do not make the generative AI tool used in the experiment

The lack of control over the development of the generative AI tool used in the experiment introduces a potential limitation. It restricts the ability to modify or customize the tool based on specific research needs and constraints. Although we choose Leonardo.AI because it provides the most complete functionality, building a tool in-house would provide more flexibility and control over its functionalities and design.

5.3 Future Work

Moving forward, several additional aspects can be considered for future work to enhance the generative AI tool and further investigate its capabilities:

Include more participants with higher diversity

Expanding the participant pool by including individuals from diverse backgrounds, such as different age groups, professional domains, and cultural backgrounds, would enhance the generalizability of the findings. A larger and more diverse participant group would provide broader perspectives and insights into the tool's usability and effectiveness across different user demographics.

Compare the group of experienced and inexperienced users

As generative tools have become more popular and broadly used, It will be possible to conduct a comparative analysis between users with varying levels of experience and expertise in using generative AI tools. By dividing users into groups based on their experience, it would be possible to explore how different user groups interact with the tool and identify areas for improvement specific to each group's needs.

Expert Evaluation

Integrating expert evaluations from experienced designers or design professionals could provide a more in-depth and objective assessment of the generated designs' quality. Expert feedback would complement user feedback and contribute to a more comprehensive evaluation of the tool's effectiveness.

Develop the generative AI tool from scratch for better customization

Building the generative AI tool in-house would provide greater control and flexibility over its functionalities, design, and integration with other systems. By developing the tool from scratch, researchers would have the opportunity to tailor it precisely to the study's objectives, ensuring that it aligns closely with the specific research requirements and can be enhanced based on ongoing feedback and insights.

Explore different product domains that are more general but not overly simplistic

Instead of focusing solely on bikes, selecting a diverse range of products from various industries would offer a broader understanding of how the generative AI tool performs in different contexts. Choosing more general but not overly simplistic products would allow researchers to evaluate the tool's applicability and performance across a wider range of design scenarios, uncovering potential strengths and limitations.

Comparative Studies

Conducting comparative studies between different generative AI tools or versions of the same tool could help evaluate their strengths and weaknesses. Comparing various tools based on user satisfaction, design quality, and ease of use could aid in identifying the most effective design tool for specific design tasks.

Incorporate eye-tracking technology to analyze users' visual attention and behavior

Integrating eye-tracking technology into the experiment would provide valuable insights into users' visual attention and behavior while interacting with the generative AI tool. By tracking users' eye movements and gaze patterns, researchers can gather data on which features, elements, or areas of the tool attract the most attention, helping identify areas for improvement in terms of visual hierarchy, user interface design, and user experience.

Combine the usage of other tools like text-to-prompt or 3D object generative AI tool

Integrating the generative AI tool with other complementary tools, such as a text-to-prompt tool or a 3D object generative AI tool, can enhance the overall design workflow and expand the creative possibilities for users. A text-to-prompt tool could allow users to input textual descriptions or concepts, which can then be converted into prompts for the generative AI tool. This integration would provide users with an additional means of communicating their design ideas and preferences to the AI model. Similarly, incorporating a 3D object generative AI tool would enable users to generate three-dimensional designs or incorporate three-dimensional elements into their designs, further expanding the tool's capabilities and enabling users to explore more complex and realistic design concepts. However, maintaining a balance between user freedom and control within the experiment setting poses a challenging problem.

By considering and implementing these suggestions in future work, researchers can gain deeper insights into its usability and performance across diverse user groups and improve the design that enhances the generative AI tool's functionality and user experience.

6 Conclusion

To answer the first research question: "How are users currently using prompt-to-image Generative AI tools to design products that meet their requirements?", we explored user approaches in designing products with prompt-to-image Generative AI tools. Through the analysis of user behavior and prompt usage, we identified several key findings:

- Users commonly employ a combination of global and local editing, with a significant focus on local refinements. They spend approximately 1/4 of their time on global editing and 3/4 on local editing, indicating a preference for refining specific details to meet their requirements.
- During global editing, users focus on improving the novelty of the generated images. They tend to allocate more time and use a higher number of prompts to explore and create novel ideas.
- In the local editing phase, users focus on improving the design's feasibility. They tend to spend more time and use more prompts to ensure the practical aspect of the generated images.
- Users naturally combine prompts that improve both feasibility and aesthetics, reflecting their understanding that design should be both visually appealing and practical.
- Common topics in participants' prompts include specific bike parts for feasibility, future technologies and innovative features for novelty, and aspects like color, shape, and pattern for aesthetics.

Through the execution of the experiment and analysis of the collected data, we aimed to answer the second research question, "What prompts and approaches yield better results?" In exploring how users utilized the prompt-to-image Generative AI tool to design bike designs that met the requirements of feasibility, novelty, and aesthetics, we identified several strategies for success.

- The interplay between user involvement and the AI's creative process in design tasks significantly impacts the outcome. Users who provide high-level,

exploratory prompts allow the AI to exercise creativity and generate unique and innovative designs. On the other hand, users who use directive, specific prompts guide the AI more precisely to produce designs that align with their preconceived ideas. Both approaches can result in aesthetically pleasing and novel designs. However, the fact that directive prompt users achieved higher feasibility ratings suggests that clear and detailed instructions help hold the AI's creativity within practical boundaries.

- Initiating evaluations with multi-criteria prompts leads to better results in the design process. By considering and communicating a broader range of design attributes, users can guide the generative AI tool in creating more comprehensive and well-rounded designs.
- There is a positive correlation between aesthetics-related prompts and both aesthetics and feasibility ratings. Considering aesthetics during the design process enhances the visual appeal and practicality of the generated designs, known as the "aesthetic-usability" or "aesthetic-utility" effect. Incorporating aesthetically pleasing elements leads to more appealing and user-satisfied designs with improved feasibility for real-world applications.
- Clarity and specificity are more important than the length of the prompts. Users should focus on providing clear, concise, and specific instructions to enhance idea generation effectively. Detailed prompts that address relevant aspects of the design are more advantageous in achieving desired outcomes than longer but less specific prompts.
- Guidelines for good prompting include specificity for feasibility, requesting innovative elements for novelty, and using expressive language for aesthetics. Considering materials, components, and manufacturing processes for feasibility, integrating futuristic technologies for novelty, and specifying desired design themes for aesthetics help generate high-quality bike designs.
- There is a trade-off between feasibility and novelty in the generated results. When feasibility increases, novelty generally decreases, and vice versa. Users should be aware of this trade-off and set realistic expectations and priorities when generating prompts to align with their design goals.

The limitations of this experiment, such as the small sample size and unfamiliarity with the generative AI tool, should be addressed in future work. Additionally, incorporating eye-tracking technology and integrating the generative AI tool with other complementary tools could enhance the overall design workflow and user experience.

In conclusion, the findings provide valuable insights into the user approaches and strategies for success when using prompt-to-image generative AI tools. By considering these insights and recommendations, researchers and designers can enhance the effectiveness and usability of such tools, and users can utilize these strategies to pursue better outcomes.

Bibliography

[Che+18] Xiang'Anthony' Chen, Ye Tao, Guanyun Wang, et al. "Forte: User-driven generative design". In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–12 (cit. on pp. 8, 9).

[DMB22] Hai Dang, Lukas Mecke, and Daniel Buschek. "GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–15 (cit. on p. 7).

[DN21] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794 (cit. on p. 6).

[DSB16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016) (cit. on p. 5).

[GL23] Nico Giambi and Giuseppe Lisanti. "Conditioning Diffusion Models via Attributes and Semantic Masks for Face Generation". In: *arXiv preprint arXiv:2306.00914* (2023) (cit. on p. 6).

[GH16] Carlos A. Gomez-Uribe and Neil Hunt. "The Netflix Recommender System: Algorithms, Business Value, and Innovation". In: *ACM Trans. Manage. Inf. Syst.* 6.4 (Dec. 2016) (cit. on p. 3).

[Goo+20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144 (cit. on p. 4).

[Gra+18] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. "Ffjord: Free-form continuous dynamics for scalable reversible generative models". In: *arXiv preprint arXiv:1810.01367* (2018) (cit. on p. 5).

[Här+20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. "Ganspace: Discovering interpretable gan controls". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9841–9850 (cit. on pp. 6, 7).

[Ho+19] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. "Flow++: Improving flow-based generative models with variational dequantization and architecture design". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2722–2730 (cit. on p. 6).

[Iso+17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134 (cit. on p. 5).

[JS91] David G Jansson and Steven M Smith. “Design fixation”. In: *Design studies* 12.1 (1991), pp. 3–11 (cit. on p. 28).

[Kar+17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. “Progressive growing of gans for improved quality, stability, and variation”. In: *arXiv preprint arXiv:1710.10196* (2017) (cit. on p. 5).

[Kaz+17] Rubaiat Habib Kazi, Tovi Grossman, Hyunmin Cheong, Ali Hashemi, and George W Fitzmaurice. “DreamSketch: Early Stage 3D Design Explorations with Sketching and Generative Design.” In: *UIST*. Vol. 14. 2017, pp. 401–414 (cit. on pp. 8, 9).

[KD18] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative flow with invertible 1x1 convolutions”. In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 6).

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. “Matrix factorization techniques for recommender systems”. In: *Computer* 42.8 (2009), pp. 30–37 (cit. on p. 4).

[KG22] Yuki Koyama and Masataka Goto. “BO as Assistant: Using Bayesian Optimization for Asynchronously Generating Design Suggestions”. In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 2022, pp. 1–14 (cit. on pp. 6, 7).

[KK95] Masaaki Kurosu and Kaori Kashimura. “Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability”. In: *Conference companion on Human factors in computing systems*. 1995, pp. 292–293 (cit. on p. 33).

[LSY03] Greg Linden, Brent Smith, and Jeremy York. “Amazon. com recommendations: Item-to-item collaborative filtering”. In: *IEEE Internet computing* 7.1 (2003), pp. 76–80 (cit. on p. 3).

[Lin+10] Julie S Linsey, Ian Tseng, Katherine Fu, et al. “A study of design fixation, its mitigation and perception in engineering design faculty”. In: (2010) (cit. on p. 28).

[Mao+17] Xudong Mao, Qing Li, Haoran Xie, et al. “Least squares generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2794–2802 (cit. on p. 5).

[Miy+18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. “Spectral normalization for generative adversarial networks”. In: *arXiv preprint arXiv:1802.05957* (2018) (cit. on p. 5).

[ND21] Alexander Quinn Nichol and Prafulla Dhariwal. “Improved denoising diffusion probabilistic models”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8162–8171 (cit. on p. 5).

[Ouy+22] Long Ouyang, Jeff Wu, Xu Jiang, et al. “Training language models to follow instructions with human feedback”. In: *arXiv preprint arXiv:2203.02155* (2022) (cit. on p. 10).

[Pen+13] David M Pennock, Eric J Horvitz, Steve Lawrence, and C Lee Giles. “Collaborative filtering by personality diagnosis: A hybrid memory-and model-based approach”. In: *arXiv preprint arXiv:1301.3885* (2013) (cit. on p. 3).

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv preprint arXiv:1511.06434* (2015) (cit. on p. 5).

[Ram+21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al. “Zero-shot text-to-image generation”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831 (cit. on p. 8).

[Ree+16] Scott Reed, Zeynep Akata, Xinchen Yan, et al. “Generative adversarial text to image synthesis”. In: *International conference on machine learning*. PMLR. 2016, pp. 1060–1069 (cit. on p. 5).

[RRS15] Francesco Ricci, Lior Rokach, and Bracha Shapira. “Recommender systems: introduction and challenges”. In: *Recommender systems handbook*. Springer, 2015, pp. 1–34 (cit. on p. 3).

[Sal+16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, et al. “Improved techniques for training gans”. In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 5).

[Shu+20] Dule Shu, James Cunningham, Gary Stump, et al. “3d design using generative adversarial networks and physics-based validation”. In: *Journal of Mechanical Design* 142.7 (2020), p. 071701 (cit. on p. 10).

[SS10] Andreas Sonderegger and Juergen Sauer. “The influence of design aesthetics in usability testing: Effects on user performance and perceived usability”. In: *Applied ergonomics* 41.3 (2010), pp. 403–410 (cit. on p. 33).

[VPT16] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. “Generating videos with scene dynamics”. In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 5).

[YM20] Chenxi Yuan and Mohsen Moghaddam. “Attribute-aware generative design with generative adversarial networks”. In: *Ieee Access* 8 (2020), pp. 190710–190721 (cit. on p. 10).

[ZB21] Enhao Zhang and Nikola Banovic. “Method for exploring generative adversarial networks (GANs) via automatically generated image galleries”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15 (cit. on pp. 7, 8).

[Zha+19] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. “Deep learning based recommender system: A survey and new perspectives”. In: *ACM Computing Surveys (CSUR)* 52.1 (2019), pp. 1–38 (cit. on p. 4).