



# **Genjourney: A Text-to-Image Prompt Optimization System Based on Interactive Genetic Algorithm and Large Language Models**

A dissertation submitted for the degree of  
**MSc Scientific and Data Intensive Computing**  
of  
**University College London**

**Fangzhou Zhang**

Department of Physics and Astronomy  
August, 2023

I, Fangzhou Zhang, confirm that the work presented in this dissertation is my own. Where information has been derived from other sources, I confirm that this has been indicated in the dissertation.

Link of the GitHub repository: <https://github.com/Arkzf2/Genjourney>

## **Abstract**

Amid the propulsion of generative artificial intelligence, text-to-image generation models have experienced significant advancements in recent years, now capable of producing high-quality images based on textual cues. Yet, crafting prompts that accurately align with user requirements and aesthetics remains an intricate challenge. In response to this, we introduce "Genjourney" — a dedicated optimization system designed specifically for AI image generation models, with an emphasis on the "Midjourney" text prompts. This system is rooted in the interactive genetic algorithm and large language models.

Harnessing the formidable capabilities of large language models in natural language processing, Genjourney systematically formulates textual prompts. Moreover, the platform synergizes users' aesthetic feedback with the genetic algorithm, adopting a human-centric approach where feedback steers natural selection and determines population fitness, fostering the evolutionary optimization of prompts. Our empirical investigation reveals that Genjourney substantially elevates the alignment of generated images with users' expectations. With merely a concise initial prompt, the system can engender results in its initial population that surpass those produced by the foundational prompts.

Furthermore, Genjourney demonstrates a keen aptitude for converging towards user-specified image objectives, with an observed augmentation in user satisfaction throughout its interactive evolutionary process. This study accentuates the potential of intertwining interactive genetic algorithms with expansive language models in the realm of Human-Computer Interaction (HCI).

## **Acknowledgement**

First and foremost, I would like to extend my profound gratitude to my research supervisor, Dr Peter Bentley, for his unwavering guidance and invaluable feedback throughout the entirety of this research journey. I am particularly appreciative of his meticulous review of my drafts, constructive suggestions, and the challenges they presented to me, urging me to think both critically and expansively, especially during the most demanding phases of this research.

Furthermore, I wish to express my heartfelt appreciation to my family for their unwavering love and understanding. Their unconditional support for my academic endeavors and their constant concern for my well-being have been pillars of strength.

Lastly, I extend my deepest gratitude to my girlfriend. Her silent encouragement and steadfast companionship during this period have been invaluable. Her significant assistance in various aspects of my life cannot be overstated.

I am thankful for all the serendipitous encounters along the way.

# Contents

1. Introduction.....	1
1.1. Background of AI Image Generator .....	1
1.2. The Importance of Accurate and Descriptive Prompts .....	1
1.3. Objective of the Study .....	1
2. Literature Review.....	2
2.1. Existing Methodologies in AI Image Generation .....	2
2.1.1. Generative Adversarial Networks (GAN).....	2
2.1.2. Variational Autoencoders (VAE).....	2
2.1.4. Diffusion Models .....	4
2.1.5. DALL-E2 vs Stable Diffusion vs Midjourney .....	5
2.2. Role of Textual Prompt and Prompt Engineering in AI Image Generation ..	5
2.3. Existing Strategies for Prompt Engineering and Optimizing Prompts.....	6
2.4. Overview of Genetic Algorithms and Their Applications .....	7
3. Methodology .....	8
3.1. Selection of Text Prompt Modifiers .....	8
3.2. Prompt Engineering Using Large Language Models (LLM) .....	9
3.2.1. Prompt Template for the Expansion of the Initial Prompt's Subject	10
3.2.2. Prompt Template for Other Six Modifiers .....	11
3.2.3. Chain Engineering .....	13
3.2.4. Natural Selection and Fitness Evaluation Based on User Feedback	13
3.2.5. Crossover .....	13
3.2.6. Mutation.....	14
3.3. Genjourney--A Web Application of Prompt Optimization System .....	15
3.3.1. Workflow of Genjourney .....	15
3.3.2. Components in Genjourney .....	16
4. Experiments and Results.....	17
4.1. Experiment on the Repetition Rate of Prompt Modifiers Generated by GPT-3.5-turbo.....	17
4.1.1. Objective .....	18
4.1.2. Setup .....	18
4.1.3. Results.....	19
4.1.4. Analysis.....	20
4.2. Comparative analysis focusing on the aesthetic differences between images generated with and without the inclusion of prompt modifiers .....	21
4.2.1. Objective .....	21
4.2.2. Setup .....	21
4.2.3. Results.....	22
4.2.4. Analysis.....	23
4.3. Assessment of the prompt optimization system's capacity to evolve towards a desired target image .....	24
4.3.1. Objective .....	24
4.3.2. Setup .....	24

4.3.3.	Results.....	25
4.3.4.	Analysis.....	32
4.3.5.	Expansion Experiment of Mutation Rate.....	33
4.4.	User Satisfaction Survey for Prompt Optimizing System.....	34
4.4.1.	Objective .....	34
4.4.2.	Setup .....	34
4.4.3.	Results.....	35
4.4.4.	Analysis.....	35
5.	Conclusion .....	36
5.1.	Summary .....	36
5.2.	Limitation .....	37
5.2.1.	Limitation on Semantic Similarity Improvement .....	37
5.2.2.	Subjectivity of Aesthetic Preferences .....	37
5.2.3.	Model Constraints.....	37
5.2.4.	Dependence on Initial Prompts.....	38
5.3.	Recommendation.....	38
5.3.1.	Utilizing More Advanced Large Language Models for Prompt Modifier Generation.....	38
5.3.2.	Fine-tuning Models for User Personalization .....	38
	References.....	38

# **1. Introduction**

In recent years, artificial intelligence (AI) has reached a developmental inflection point. Particularly in the realm of artificial intelligence-generated content (AIGC), there has been a significant enhancement in its capabilities [4]. This has enabled AI to not only demonstrate elevated levels of creativity but also to achieve remarkable advancements in aesthetic design, thereby establishing it as a pivotal technological tool in the digital age.

## **1.1. Background of AI Image Generator**

The advancement of artificial intelligence in image generation is closely tied to the progress of deep learning technologies. As early as the 1970s, artist Harold Cohen began developing the "ARRON" program to facilitate computer-aided painting [48]. By 2014, Ian Goodfellow introduced the Generative Adversarial Network model (GAN), marking a significant leap in the evolution of AI-driven art [11]. Within this model, a generator is tasked with producing images, while a discriminator assesses the quality of those images, the two working synergistically. In 2021, OpenAI released the deep learning model CLIP, considered one of the most advanced image classification models of its time [28]. Building upon these foundations, researchers refined the diffusion model, leading to the introduction of the stable diffusion model, now regarded as one of the foremost AI painting models in the field [39].

## **1.2. The Importance of Accurate and Descriptive Prompts**

Text-to-image generative models, such as Stable Diffusion, DALL-E2 and Midjourney, have the capability to produce high-quality images based on natural language descriptions, commonly referred to as prompts [19, 37, 39]. These generated images find applications in a broad spectrum of domains including digital art, industrial design, and craft education [47, 51]. The ability to precisely control image generation through prompts is of paramount importance. However, mastering this capability remains a significant challenge [16]. Ambiguous or inaccurate prompts may lead to the model producing images that diverge from expectations, whereas clear and precise prompts can notably enhance the aesthetic and visual appeal of images produced by AI models [49].

## **1.3. Objective of the Study**

Given the central role of textual prompts in AI-driven image generation, this study is dedicated to the optimization of these prompts with an aim to enhance the aesthetic and visual quality of the resultant images. Utilizing Large Language Model, we will refine prompts to encompass key elements such as artistic style, color, lighting effects, and rendering quality. Using an interactive genetic algorithm, we iteratively explored textual prompt and effectively approached the images most satisfactory to users by

collecting their rating feedback. By synergizing user preferences with the proposed prompt optimization techniques, we aspire to elevate the image generation process from a user-centric aesthetic perspective.

## **2. Literature Review**

In recent years, technological breakthroughs in the field of artificial intelligence, particularly advancements in Artificial Intelligence Generated Content (AIGC), have piqued the interest of numerous scholars [7, 42]. This progress is evident not only in algorithmic optimizations but also in the broad and intricate real-world applications.

To delve deeper into this research domain, we undertook a critical review of relevant literature. Our focus was primarily on the following aspects: existing methodologies in AI image generation, contemporary image-generating models, the role of textual prompts and prompt engineering in AI image generation, some existing strategies for optimizing prompts and overview of genetic algorithms with their applications.

### **2.1. Existing Methodologies in AI Image Generation**

#### **2.1.1. Generative Adversarial Networks (GAN)**

Generative Adversarial Networks (GANs) are an image generation approach based on deep learning technologies, initially proposed by Goodfellow et al. in 2014 [11].

This innovative framework consists of two synergistically operating neural networks: a generator and a discriminator. The essence of GANs lies in the generator creating images with random noise vectors as input, refining its output iteratively to deceive the discriminator. The discriminator's role is to differentiate between real images (sourced from datasets) and the fabricated images produced by the generator. This adversarial game, inspired by the zero-sum game concept in game theory, ensures that both neural networks enhance their performance iteratively, leading to the generator producing increasingly realistic images and the discriminator becoming more adept at discerning their authenticity.

Although GANs do have certain limitations, such as a restricted variety of images generated by the generator and unstable training, it's undeniable that GANs have marked a groundbreaking innovation in generative models.

#### **2.1.2. Variational Autoencoders (VAE)**

Variational Autoencoders (VAE) stand as a cornerstone technique in the realm of generative models. Introduced by Kingma & Welling in 2013, VAEs not only imparted a probabilistic viewpoint to the classical autoencoders but also bridged the gap between deep learning and Bayesian inference [14]. Essentially comprised of an encoder and a



decoder, the encoder maps the input data to a latent space, outputting a mean and a standard deviation of a Gaussian distribution. Subsequently, the decoder samples from this distribution and reconstructs the original data based on these latent variables. This sampling mechanism infuses the model with stochasticity, aiding in the generation of diverse new samples.

VAEs have found applications in various domains including image generation, denoising, and recently in generating non-image data [15]. Owing to their capability to create continuous latent spaces, VAEs showcase unique advantages in tasks like image interpolation and morphing [34].

However, it is pertinent to note that while VAEs have multiple merits, they often produce images that are somewhat blurrier compared to those generated by GANs, especially when dealing with high-resolution and diverse image datasets.

### **2.1.3. Contrastive Language–Image Pre-training (CLIP)**

OpenAI's CLIP (Contrastive Language–Image Pre-training) represents a pioneering approach in the realm of AI image generation and recognition. Introduced in 2021, it symbolizes a paradigm shift away from traditional training methodologies [30]. In contrast to typical visual models that are trained on specific datasets for singular tasks, CLIP is designed to understand images in tandem with natural language, facilitating its capability to tackle a variety of visual tasks without task-specific training.

The essence of CLIP is rooted in its unique training scheme. It learns from a vast collection of internet text-image pairs [30]. In doing so, the model is adept at associating images with an array of linguistic descriptions, bypassing the reliance on class labels commonly found in conventional datasets. This training paradigm, anchored in contrastive learning, equips the model to match images with textual descriptions and vice versa.

The versatility of CLIP is evident in its ability to execute a multitude of tasks, such as zero-shot image classification, geolocation, and even creating sketches based on textual prompts. Its performance often rivals, if not surpasses, traditional models that have been meticulously trained for specific purposes.

However, like any model, CLIP is not devoid of limitations. Potential biases from its training dataset or challenges in addressing ambiguous prompts can be areas of concern. Nonetheless, its introduction undeniably signifies a notable advancement in the evolution of visual models, laying the foundation for more integrated and generalized AI systems.

The DALL-E2 model, released by OpenAI in 2022, effectively utilizes the CLIP model to achieve innovative performance in generating images from text [37]. The DALL-E2 model comprises three primary components: CLIP, a prior module, and an image decoder. During the model's training phase, these sub-modules are trained

independently [27]. Once individually trained, these modules are integrated to collaboratively achieve the transformation from text to image.

In the practical application of the model, the image encoder section of the CLIP module is discarded, retaining only the text encoder. This is done because the objective is to generate images from text, necessitating only the text encoding part. Subsequently, the text encoder encodes the text, followed by the prior module which further translates this text encoding into image-associated encodings. Finally, the image decoder is responsible for converting these encodings into specific image outputs.

#### **2.1.4. Diffusion Models**

Diffusion models introduce a novel paradigm in the domain of AI image generation. Contrary to conventional generative models, diffusion models adopt a reverse approach. Starting with the target image, they progressively introduce noise in a controlled manner until the image is transformed into pure noise. The generative procedure then involves reversing this noise addition to reconstruct the original image [51].

Inspired by the principles of diffusion physics and stochastic processes, these models aim to deeply understand the dynamics of data transformations. Given an initial data point, diffusion models can predict its temporal evolution under the influence of random "noise." In the realm of imagery, this evolution is perceived as a transition from a coherent image to a random noise pattern and vice versa.

Diffusion models have demonstrated significant potential in image generation. Their uniqueness lies in the capability of the model to grasp the entire trajectory of the image transitioning from clarity to noise and back, allowing it to generate detailed imagery nuances that might be challenging for other generative models. Moreover, due to their incremental noise addition trait, they are inherently more robust against issues like mode collapse, commonly faced by models like GANs.

The emergence of diffusion models has ushered in innovative breakthroughs in the realm of AI-generated art. A recent advancement, the Stabilized Diffusion Model, also known as the Latent Diffusion Model (LDM), offers an efficient tool for crafting artistic illustrations, enabling users to create them through simple textual prompts. Distinguished from traditional diffusion models, the primary differentiation of LDM lies in its diffusion process occurring within a latent space, which accelerates its image generation speed.

Beyond the Stable Diffusion model, Midjourney represents another groundbreaking AI art image generation model based on diffusion principles. Distinguishing itself from other diffusion models, Midjourney is trained in tandem with large language models, bestowing it with a unique image generation capacity [40]. This model excels in manipulating and amalgamating foundational artistic styles, producing distinct images

tailored to user preferences. Particularly noteworthy is Midjourney's proficiency in crafting immersive environments, especially in the realms of fantasy and science fiction. The dramatic lighting effects it employs infuse its images with depth and vivacity, reminiscent of conceptual art extracted from high-end video games.

### 2.1.5. DALL-E2 vs Stable Diffusion vs Midjourney

Given the primary goal of this study to augment the aesthetic appeal of generated images via optimized text prompts, it's imperative to adopt a model that consistently delivers top-tier visual output. After rigorously evaluating the image quality rendered by the previously discussed models (DALL-E2, Stable Diffusion, and Midjourney) based on various parameters such as texture detail, color coordination, and overall visual engagement, Midjourney emerged as a clear front-runner [5]. Consequently, recognizing its unmatched prowess in generating visually arresting imagery as shown in Table 2.1, Midjourney was selected as the AI image-generation tool for our research.










Prompt	DALL-E2	Stable Diffusion	Midjourney
“a cyberpunk city”			
“a day in the life of a robot”			
“a heart-shaped sculpture made of crystal”			

Table 2.1: Images generated by DALL-E2, Stable Diffusion, and Midjourney

## 2.2. Role of Textual Prompt and Prompt Engineering in AI

## Image Generation

In the domain of AI image generation, textual prompts have increasingly become the pivotal driver. The state-of-the-art image generation models predominantly rely on textual cues to produce corresponding images. Consequently, the quality of these textual prompts directly determines the outcome of the generated imagery. A groundbreaking study by Reed et al. in 2016 underscored the paramount significance of detailed textual descriptions in creating relevant images [38]. Their model, conditioned on textual descriptions, notably excelled in generating images of birds and flowers, exhibiting remarkable results. With the advent of models like DALL-E2, Stable Diffusion, and Midjourney, the importance of prompts has been further accentuated [19, 37, 39]. Intriguingly, slight tweaks in the prompts can lead to profound disparities in the visual aesthetics of the produced images. As AI image generation evolves, especially with a trend towards larger models, an ambiguous prompt may not fully harness the potential of these powerful AI generation capabilities.

The concept of "prompt engineering" was originally introduced by Gwern Branwen while devising text inputs for OpenAI's GPT-3 language model [6]. Here, the term "engineering" doesn't refer to the stringent sciences commonly associated with the STEM fields. Instead, prompt engineering delves into a practice immersed in natural language processing, entailing the manual crafting and tweaking of prompts to hone model outputs. Consequently, its nature aligns more with human-computer interaction than the conventional domains of machine learning.

Within the flourishing domain of AI-infused digital art, prompt engineering has established itself as a pivotal component. Owing to the inherent potential of AI image-generating models, varying prompt keywords can elicit drastically diverse aesthetic results, toggling between "surrealist style" "Disney style," and "abstract style". Given the expansive datasets these AI tools are trained on, a dual-edged sword emerges. The breadth of training material assures varied responses, yet many users struggle to encapsulate their imaginative vision due to a vocabulary deficit specific to a particular art style. In light of this, the conceptualization of an automated prompt optimization system stands out as a prospective trajectory in the realm of prompt engineering research [33].

### 2.3. Existing Strategies for Prompt Engineering and Optimizing

#### Prompts

Oppenlaender undertook a comprehensive investigation into prompt engineering with a specific emphasis on its role in text-to-image synthesis [32]. A cornerstone of this research centered on the ability of users to modulate model outcomes by integrating distinct key phrases, termed as "modifiers." Spanning an exhaustive three-month ethnographic exploration, Oppenlaender identified six unique categories of prompt

modifiers: Subject term, Style modifier, Image prompt, Quality booster, Repeating term, and Magic term. These delineations were crafted to bestow users with granular control over image outputs, ensuring alignment with their creative aspirations.

While the genesis of Oppenlaender's classifications originates from intricate analyses within online communities and introspective ethnographic studies, their significance is paramount. Recognizing that the domain of prompt engineering, particularly when contextualized within Human-Computer Interaction (HCI), is embryonic, such pioneering contributions become instrumental. They not only offer invaluable insights but also harbor the potential for broad-based applications in future research trajectories. The structuring of prompts within this study leans heavily on the modifier typologies proposed in Oppenlaender's influential contribution.

Ustalov and Pavlichenko pioneered an approach within human-computer interaction, utilizing genetic algorithms to learn and refine combinations of prompt keywords to bolster the aesthetic quality of generated images [35]. Through an analysis of the top 100 keywords from the widely-engaged Stable Diffusion Discord and leveraging genetic algorithms, they optimized the keyword set based on its average ranking. The results were commendable; their algorithmically-refined keyword set outperformed the top 15 keywords. Moreover, their findings underscored the significance of prompt keywords: the presence of any prompt keyword invariably enhanced image quality in contrast to its absence.

While this research elucidates an optimized keyword set that can potentially augment the caliber of prompt-generated images, it isn't without limitations. The constraint to only 100 keywords might inadvertently omit numerous other relevant keywords. Furthermore, this methodology might be tinged with "community bias", given its reliance on preferences exclusive to the Stable Diffusion Discord community. One must acknowledge the intrinsic subjectivity embedded in aesthetic judgments. The perception of beauty and allure is deeply individualistic. Therefore, in the pursuit of a genuinely tailored prompt optimization system, the chosen prompts should resonate with the distinct stylistic inclinations and preferences of individual users.

## **2.4. Overview of Genetic Algorithms and Their Applications**

The Genetic Algorithm (GA) represents a search and optimization technique inspired by the mechanisms of natural selection, and it has been widely employed within the artificial intelligence domain [12]. The foundational principle of the Genetic Algorithm encompasses three stages: selection, crossover (or recombination), and mutation, thereby emulating the biological evolutionary process [1]. Due to its prowess in discerning optimal solutions in diverse intricate scenarios, the Genetic Algorithm manifests tremendous potential in AI research [3]. In the context of this study, we adopt Interactive Genetic Algorithm (IGA), where fitness is determined by users [17]. Owing

to its inherent characteristics, IGA offers a solution to challenges that are not readily addressed by GA, including those in design and art [26]. This facilitates the crossover and mutation of textual prompts, simulating an evolutionary process, aiming to refine the quality of the generated images.

### 3. Methodology

The central objective of this research is to design and validate a system for optimizing textual prompts in a text-to-image AI model. Our approach involves the selection of textual keyword elements, the prompt engineering using large language models, an interactive genetic algorithm mechanism grounded in user feedback and an instantiated system--Genjourney. In this study, we selected GPT-3.5-turbo as the large language model for generating textual prompt modifiers and Midjourney as the text-to-image image generation model.

#### 3.1. Selection of Text Prompt Modifiers

The textual prompting mechanism of Midjourney operates using concise textual phrases to drive AI image generation [21]. While every AI image generation model typically possesses its inherent style, utilizing a brief prompt, such as "lion," often results in generic, undistinguished image outputs, as illustrated in Figure 3.1. To achieve more unique and tailored results, it's advisable to employ more detailed and descriptive prompts, which can guide the model to produce distinct outputs.






**Figure 3.1: Images generated by the prompt “lion”**

When crafting these prompts, a meticulous consideration of every detail is imperative since any non-specified part might be randomly filled by the Midjourney model. Jeanne David (2023) underscored that for optimal generation, a structured and highly descriptive text prompt should encompass elements like subject, setting, artistic style, lighting effects, hue, and camera angles [10]. Additionally, as evidenced by the data on the Prompt Hero website, rendering is a pivotal component in Midjourney's prompting, with approximately 9000 related prompts associating with it [36]; diverse artistic styles can be achieved by integrating varying rendering effect keywords<sup>2</sup>. Thus, synthesizing the exemplary digital images and their respective prompts displayed on the Midjourney Showcase [22], this study adopts a unified text prompt framework with 7 different types of modifiers comprising the expansion of the initial prompt's subject, artistic style, artist,



color palette, composition perspective, lighting effect, rendering quality. The following Table 3.1 presents a "lion" prompt based on the textual prompt structure proposed in this study.

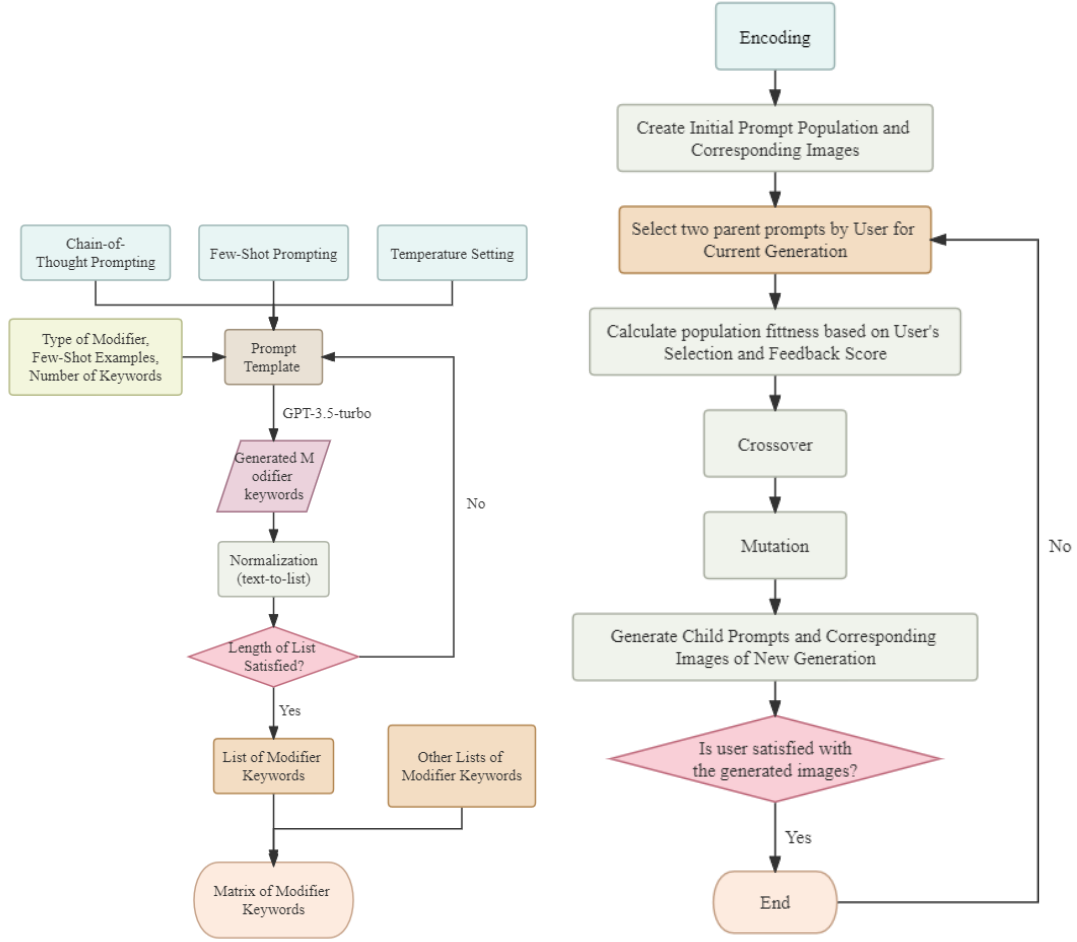
		
<p>“lion, gentle, connection, humans, bond, trust, Graffiti-inspired, Takashi Murakami, Deep mahogany, Top-down perspective, Warm sunlight, Soft shadows”</p>	<p>“lion, fierce, hunting, stealthy, moonlit forest, shadows dancing, Cubist lion, Claude Monet, Serengeti ochre, Lion from a product viewpoint showcasing its strength and beauty, Rainbow-hued halo, Texture mapping”</p>	<p>“lion, mysterious, ancient, guarding, hidden cave, glowing crystals, Futuristic lion, Georgia O’Keeffe, Simba gold, Top-down perspective of a lion on the hunt, Harsh sunlight, Specular highlights”</p>

**Table 3.1: Images generated by prompts that follows prompt framework with 7 modifiers: the expansion of the initial prompt’s subject, artistic style, artist, color palette, composition perspective, lighting effect, rendering quality**

### 3.2. Prompt Engineering Using Large Language Models (LLM)

In this study, we employed the GPT-3.5-turbo as our primary large language model. While the GPT series has already introduced the GPT-4 version, we opted for the GPT-3.5-turbo due to the unavailability of the GPT-4 API to the public at the time of our research [29].

To generate prompts that adhere to the structure proposed in this research—specifically, <expansion of the initial prompt’s subject>, <artistic style>, <artist>, <color palette>, <composition perspective>, <lighting effect>, <rendering quality>. This structure is designed to allow for precise control over image generation, enabling users to construct visual content more purposefully. To achieve optimal generation outcomes, it is essential to maintain both the diversity of generated prompts and structural standardization. To this end, we introduce a “Chain Prompt Engineering” method specifically tailored for the textual prompts in this study. The detailed flow of this method is illustrated in the left flowchart of Figure 3.2.



**Figure 3.2: Flowchart of chain prompt engineering and interactive genetic algorithm for prompt optimizing**

### 3.2.1. Prompt Template for the Expansion of the Initial Prompt's

#### Subject

The purpose of expanding upon the initial prompt subject is to provide more detailed and specific descriptions, thereby enabling the generation of images that are more accurate and align with user expectations. According to the official documentation for the Midjourney model, it is necessary to provide the GPT model with sufficient information to ensure that the text prompts it generates can be accurately understood and effectively utilized by the Midjourney model [23]. Midjourney is an efficient image generation model, capable of creating high-quality images based on detailed and specific text prompts.

In our research, we devised a prompt template, as illustrated in the following figure, based on the Persona Pattern and the Chain-Thinking approach to prompt writing. We integrated the initial prompt and the designated number of generated keywords as parameters into the template in Figure 3.3 using chain-of-thought prompting and



persona pattern [25].

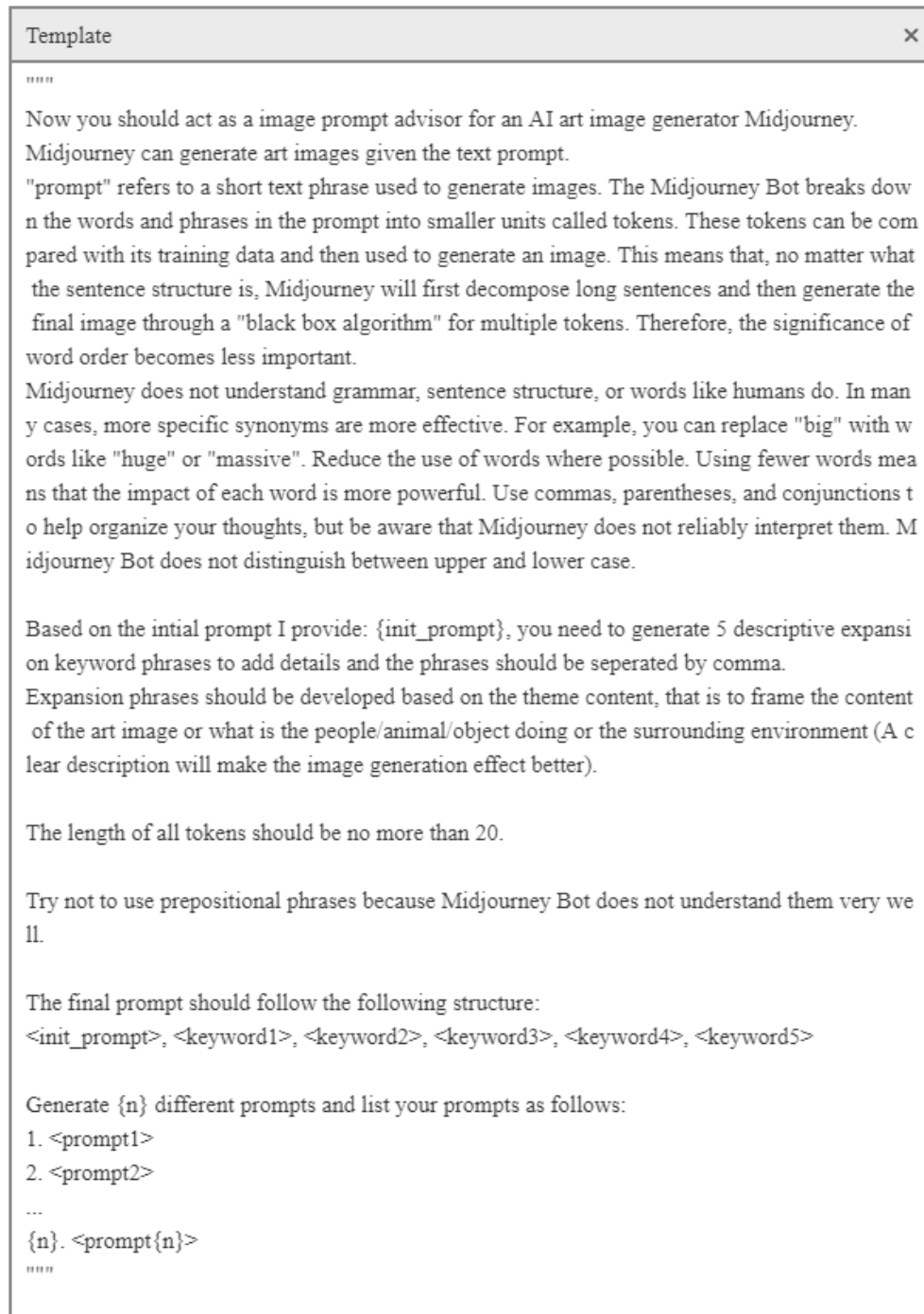
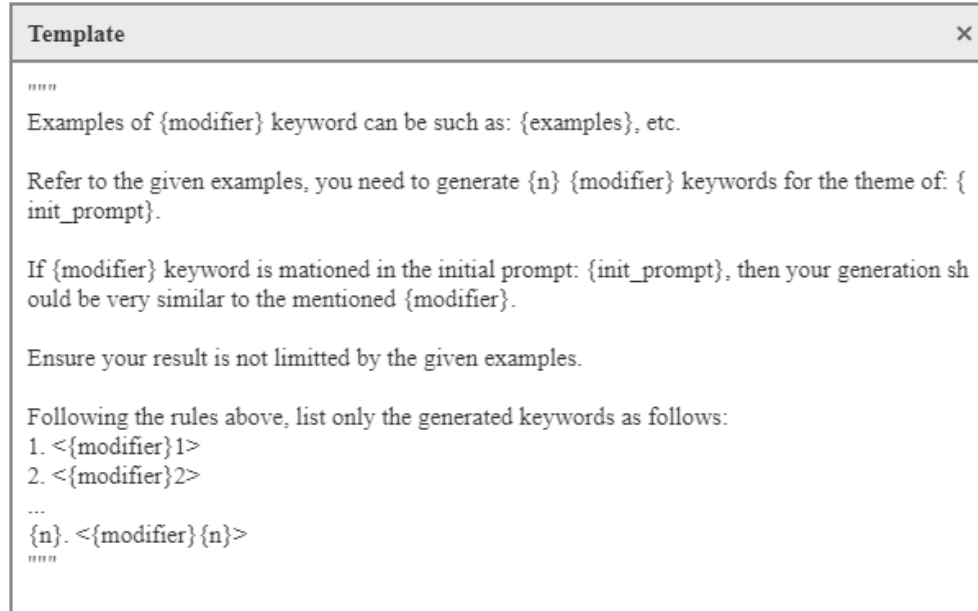


Figure 3.3: Prompt template for the expansion of the initial prompt's subject

### 3.2.2. Prompt Template for Other Six Modifiers

For the other six prompt structure components--<artistic style>, <artist>, <color palette>, <composition perspective>, <lighting effect>, <rendering quality>, we utilized a chain-of-thought approach to prompting along with a few-shot prompting

technique to draft a prompt template that serves as a runtime-instantiable parameterized prompt [25]. In the template shown in Figure 3.4, the initial prompt, prompt modifier types, few-shot examples, and the number of generated prompt keywords are integrated as parameters.



**Figure 3.4: Prompt Template for Other Six Modifiers**

For each prompt modifier, we provide ten examples as part of a few-shot approach. These examples are deliberately diverse to prevent the model from over-interpreting any specific instance. The specific examples can be found in Table 3.2.

Modifier	Examples
Artistic Style	Cyberpunk, Rococo, Futuristic, Glitch Art, Anime, Ukiyo-e, Ink painting, Minimalist, Photoreal, Pixar.
Artist	Van Gogh, Wu Guanzhong, Alphonse Mucha, Rose Tran, Jon Klassen, WLOP, Miyazaki Hayao, Jeffrey Catherine Jones, Charlie Bowater.
Color Palette	Mint Green, Pop Art, Macarons, Black and White, Soft Pink, Crystal blue, Neon Shades, Muted Tones, Maple Red, Rich Color.
Composition Perspective	Depth of Field (DOF), Long Shots, Close-up, Dynamic Symmetry, Cinematic Shot, Wide-angle view, Bird View, Golden Ratio, S-shaped Composition, Fisheye Lens.
Lighting Effect	Intense Backlight, Soft Lighting, Soft Moon Light, Studio Lighting, Crepuscular Ray, Volumetric Lighting, Front Lighting, Hard Lighting, Rainbow Halo, Glow in the Dark.

Rendering Quality	Subpixel Sampling, Arnold Renderer, V-ray Renderer, C4D renderer, Unreal Engine, Blender renderer, 4k, 3DCG, Octane Renderer, Architectural Visualization.
-------------------	--

**Table 3.2: Examples for other six modifiers**

### 3.2.3. Chain Engineering

For each type of prompt modifier, parameters were fed into the template to instantiate GPT prompts, producing 30 sets of different prompt modifiers. To ensure the diversity of generated prompt modifiers, the model's temperature was set to 0.8 [2].

Given that the GPT model essentially generates a text continuation based on probability, our current technology doesn't allow for strict output formatting solely through the language model itself [18, 46]. As such, external methods are required to standardize the model's outputs. In our research, since the model's output is in the form of a string, we utilized Python to convert this string into a list. Subsequently, we assessed whether the length of this list matched our intended count of prompt modifier sets. If not, the generation process was repeated until the specified format was achieved.

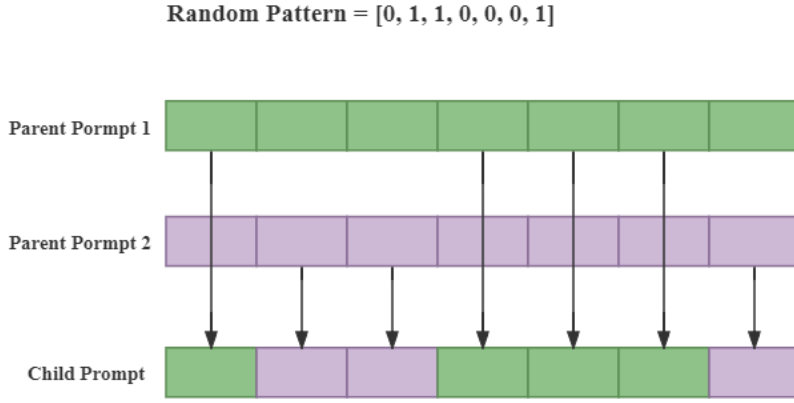
Ultimately, these lists, each with a length of 30, were amalgamated into a 30x7 text prompt matrix. For each segment of the prompt structure, modifier sets could be extracted from the matrix and combined, resulting in a vast array of potential text prompts, with  $30^7$  possible combinations. This approach not only guaranteed the diversity of prompts but also laid the groundwork for subsequent encoding methods in our genetic algorithm.

### 3.2.4. Natural Selection and Fitness Evaluation Based on User Feedback

In light of our research aim to cater to users' aesthetic preferences, we opted for a method predicated on user scores to ascertain this fitness. Considering the seven elements of the prompt structure, we compartmentalized the scoring system into five aesthetic dimensions: subject, artistic style, color palette, composition perspective and lighting effect. Herein, the dimensions of artistic style, artist, and rendering quality were integrated under the single category of artistic style. Users are tasked with selecting their top two preferred images, assigning scores to their corresponding prompts. These two prompts are then designated as parent prompts within the genetic algorithm framework. Subsequently, users rate their chosen images across the five aforementioned dimensions, with a maximum achievable score of 10 for each. Lastly, the user scores are input into the genetic algorithm as a five-element list, serving as the fitness metric for the population.

### 3.2.5. Crossover

Given that our prompts comprise multiple distinct and independent elements, we employed a random crossover method shown in Figure 3.5. within the genetic algorithm framework. For prompts composed of seven elements, we initiate the crossover by generating a random binary sequence of length seven, which we refer to as a "Random Pattern". Assuming we have two parent prompts: parent prompt 1 and parent prompt 2, a "0" in the sequence denotes the selection of a modifier element from parent prompt 1, whereas a "1" indicates a choice from parent prompt 2. As an illustrative example, let us consider a randomly generated sequence of [0, 1, 1, 0, 0, 0, 1]. This signifies that the offspring prompt will inherit the 1st, 4th, 5th, and 6th modifiers from parent prompt 1 and the 2nd, 3rd, and 7th modifiers from parent prompt 2.



**Figure 3.5: Random crossover illustration**

### 3.2.6. Mutation

In this research, the mutation rate for each gene (text prompt) is determined by the population fitness derived from user ratings. The formula for the mutation rate is shown in Equation 3.1.

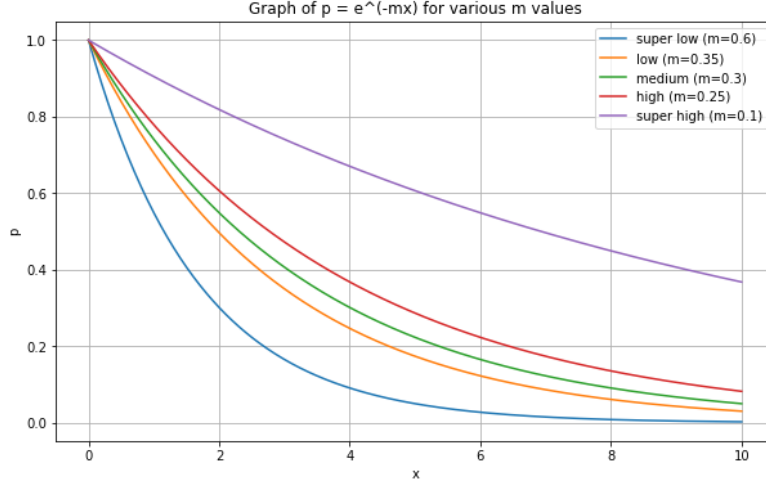
$$P_{E_i} = e^{-m \cdot s_{E_i}}$$

**Equation 3.1**

Where  $P_{E_i}$  denote the mutation rate for the  $i^{th}$  dimension out of the five dimensions.

These dimensions are: subject, artistic style, color palette, composition perspective and lighting effect which mentioned in 3.3.3. Here,  $e$  is the Euler number,  $m$  is the mutation coefficient, and  $s_{E_i}$  is the score that the user assigns to the  $i^{th}$  dimension.

In this study, we defined five mutation rates. Specifically, 'super low' corresponds to  $m = 0.6$ , 'low' is represented by  $m = 0.35$ , 'medium' is set at  $m = 0.3$ , 'high' is  $m = 0.25$ , and 'super high' is defined as  $m = 0.1$ . It's worth noting that only the 'super low' and 'super high' rates were employed in subsequent experiments. Their graphical representation within Equation 3.1 can be viewed in Figure 3.6.



**Figure 3.6: Graph of different mutation rates**

As demonstrated by the formula, a higher user score results in a lower mutation rate (close to but not equal to 0%). If a user's score is 0, the mutation rate for that element stands at 100%. Catering to diverse user requirements, the coefficients within the mutation rate formula can be adjusted, reflecting individualized aesthetic preferences. Usually, we set the mutation coefficient as 0.3.

As previously mentioned, upon user feedback, a list of length 5, with each element ranging between 0 and 10, is generated to represent the population's fitness. Based on the mutation rate formula that factors in this fitness measure, we can deduce the mutation rate for each modifier in the child prompt following crossover. Specifically, when mutation occurs within the "artistic style" aesthetic dimension, there's an equal probability of mutation for the prompt modifiers: "artistic style", "artist", and "rendering quality". The number of mutating modifiers can range from one to three. Meanwhile, mutations in the other four dimensions directly correspond to the variations in the remaining four types of prompt modifiers. Should the child prompt undergo mutation, we extract alternative modifiers from the previously established 30x7 prompt matrix, substituting the original element with a different newly selected random modifier from the respective column.

### 3.3. Genjourney--A Web Application of Prompt Optimization

#### System

In this study, we introduced a method of optimizing textual prompts via the Interactive Genetic Algorithm (IGA). Leveraging the Streamlit library, GPT-3.4-turbo API, and the Midjourney API, we developed a web application named Genjourney to instantiate this prompt optimization system [29, 43, 45].

#### 3.3.1. Workflow of Genjourney

The workflow of this system is delineated as follows:

- a) The user provides an initial prompt.
- b) Based on the initial prompt, the system formulates a prompt modifier matrix with potential combinations amounting to  $30^7$ .
- c) The system randomly constructs an initial population of structured prompts, with a population size of five, from the prompt modifier matrix.
- d) The user then selects two prompts corresponding to the images that best align with their aesthetic preferences from the current population, offering feedback from five aesthetic dimensions.
- e) Using user feedback as a measure, the system determines the fitness of the current population. Subsequent crossover and mutation operations are performed to produce the next generation of prompts.

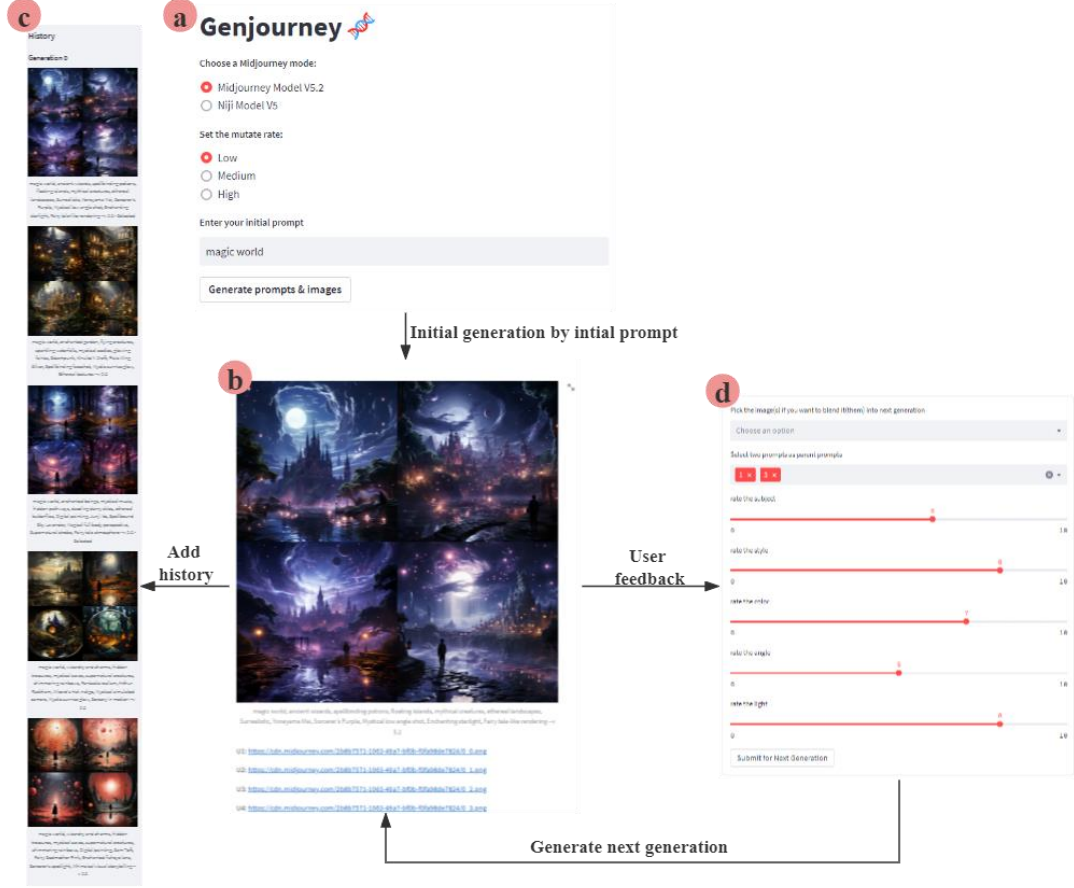
Steps d and step e are repeatedly undertaken until the resultant image aligns with the user's aesthetic criteria.

Moreover, to enhance the user's creative experience, we integrated the 'blend' feature from Midjourney, allowing users to amalgamate desirable images from the offspring with textual prompts [20].

### 3.3.2. Components in Genjourney

The Genjourney application, grounded on our proposed prompt optimization algorithm, encompasses components as depicted in Figure 3.7:

- **Model Selection:** Within Genjourney, users have access to two of Midjourney's latest models: “Midjourney Model V5.2” and “Niji Model V5” [24].
- **Mutation Rate Mode Selection:** Depending on their requirements, users can opt between three mutation rate modes: “low,” “medium,” and “high”.
- **Initial Prompt Input:** Users can articulate their concepts by feeding them into Genjourney as initial prompts.
- **Current Generation Display:** The primary interface showcases the present generation of prompt populations along with their corresponding images. Additionally, each image set is accompanied by links to its sub-images for user download convenience.
- **User-driven Interactive Genetics:** Following each population generation, users can pinpoint the prompt of the image that best resonates with them. This selected prompts serve as the parent prompts. Users then provide aesthetic feedback based on the five dimensions delineated in Section 3.2.4. Genjourney then translates this into a fitness measure for the population, perpetuating its evolution.
- **Blend Feature:** Should users desire, they can meld the prompts of up to three images from the current generation with the chosen parent prompts [20].
- **Historical Log:** All previously generated offspring populations are catalogued under the 'History' section on the page's left, with indicators specifying if they were selected.



**Figure 3.7: Workflow of Genjourney (a. Model selection, mutation rate Mode selection and input of initial prompt. b. Current generation with prompts, images and URL of sub-images. c. History log. d. User-driven interactive genetics with selection and feedback score.)**

## 4. Experiments and Results

In this section, we aim to evaluate the efficacy of the aforementioned keyword optimization system through a series of experiments. These experiments are designed to ascertain both the optimization impact on text prompts and the aesthetic enhancement of the images generated from the user's perspective. Four specific tests have been devised:

- Experiment on the repetition rate of prompt modifiers generated by GPT-3.5-turbo.
- Comparative analysis focusing on the aesthetic differences between images generated with and without the inclusion of prompt modifiers.
- Assessment of the prompt optimization system's capacity to evolve towards a desired target image.
- User satisfaction survey.

### 4.1. Experiment on the Repetition Rate of Prompt Modifiers

## Generated by GPT-3.5-turbo

### 4.1.1. Objective

The primary aim of this experiment is to investigate the diversity issue in the generation of prompt modifier keywords by the GPT-3.5-turbo model. Specifically, the research seeks to study the uniqueness ratio of prompt modifier keywords generated from both the same and different initial prompts.

### 4.1.2. Setup

Setup for the experiments to study the uniqueness ratio from both the same and different initial prompts is shown in Algorithm 4.1 and Algorithm 4.2.

---

#### **Setup for Diversity Evaluation of Prompt Modifier Keywords for a Same Initial Prompt**

---

**Input:** 10 different initial prompts  
**for** modifier in 7 modifiers **do**  
    **for** initial prompt of 10 different initial prompts **do**  
        **for** i **from** 1 **to** 10 **do**  
            generate a list of 30 modifier keywords  
        **end for**  
        calculate the uniqueness rate for all 300 generated modifier keywords  
        (=number of unique keywords / 300)  
    **end for**  
    plot the graph of uniqueness rate of 10 different initial prompts  
**end for**

---

**Algorithm 4.1: Setup for diversity evaluation for a same initial prompt**

---

---

#### **Setup for Diversity Evaluation of Prompt Modifier Keywords for Different Initial Prompts**

---

**Input:** 10 different initial prompts  
**for** modifier in 7 modifiers **do**  
    **for** i **from** 1 **to** 10 **do**  
        **for** initial prompt of 10 different initial prompts **do**  
            generate a list of 30 modifier keywords  
        **end for**  
        calculate the uniqueness rate for all 300 generated modifier keywords  
        (=number of unique keywords / 300)  
    **end for**  
    calculate the average uniqueness rate  
**end for**

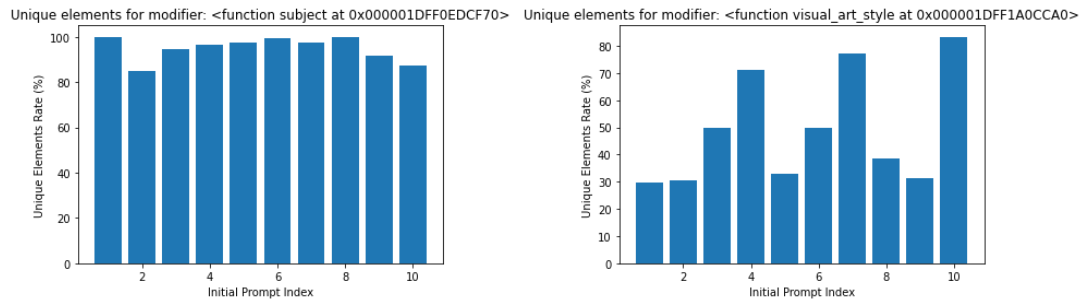
---

**Algorithm 4.2: Setup for diversity evaluation for different initial prompts**

---

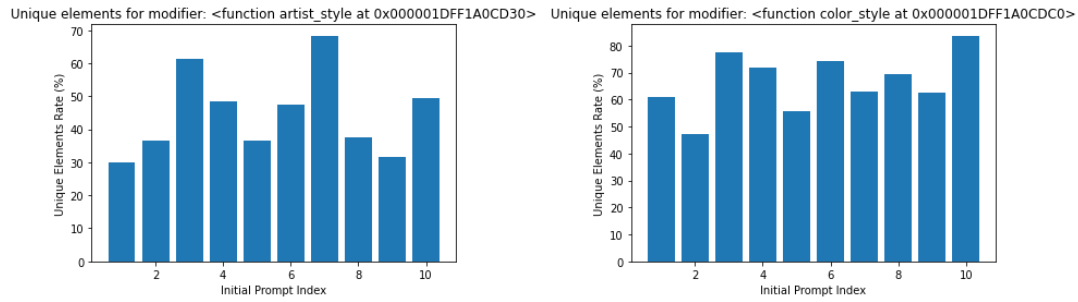


### 4.1.3. Results



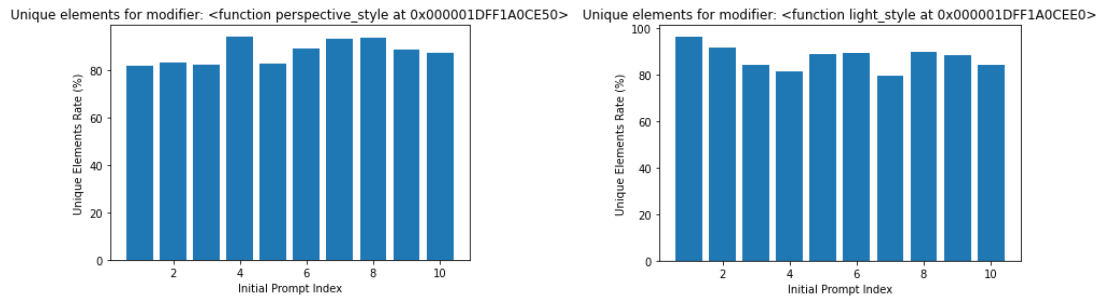
**Figure 4.1**

**Figure 4.2**



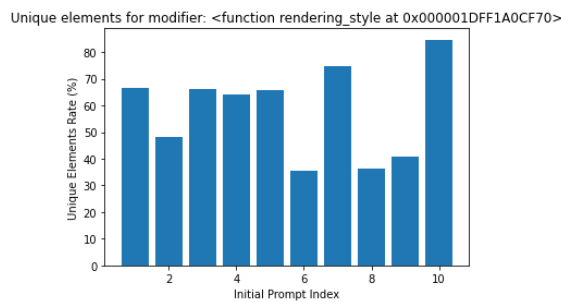
**Figure 4.3**

**Figure 4.4**



**Figure 4.5**

**Figure 4.6**



**Figure 4.7**

**Figure 4.1-4.7: Uniqueness rate for a same prompt of the prompt modifiers: expansion of the initial prompt's subject, artistic style, artist, color palette, composition perspective, lighting effect and rendering quality respectively**

	Subject	Artistic Style	Artist	Color	Perspective	Lighting	Rendering
--	---------	----------------	--------	-------	-------------	----------	-----------

Same prompt	94.9%	49.3%	44.7%	66.6%	87.7%	87.3%	58.3%
Ten different prompts	97.2%	68.6%	65.9%	96.5%	99.7%	97.8%	68.3%

**Table 4.1: Uniqueness rate of 7 modifier types for a same prompt and different prompts.**

The table and figures indicate that for the same initial prompt, repeated ten times, the uniqueness rate for five of seven types of prompt modifier exceeds 50%. Specifically, "subject," "composition perspective," and "lighting effect" categories have a uniqueness rate over 85%. The overall uniqueness rate for prompt modifier keywords is approximately 69.8%. When varying the initial prompts, the uniqueness rates for "subject," "color palette," "composition perspective," and "lighting effect" all surpass 95%, with an average uniqueness rate of 84.9%.

#### 4.1.4. Analysis

The data suggests that the GPT-3.5-turbo model, even when presented with a consistent initial prompt, can produce significant variations in the output using our prompt optimization system. For instance, as indicated in Table 4.1, the "subject" category consistently demonstrated a high uniqueness rate exceeding 90%, highlighting the model's capability in this regard.

When the model is provided with different initial prompts, it exhibits an even broader diversity, suggesting that the system is adaptive and can generate a varied set of prompt modifiers tailored to the specific subject. For example, the uniqueness rates for "subject," "color palette," "composition perspective," and "lighting effect" categories all surpass 95%, emphasizing the system's versatility.

A potential reason for the greater diversity in prompt modifier keywords generated from different initial prompts, as compared to a consistent one, could be attributed to our engineering of the prompt template (refer to Figure 4.8). In our template design, we emphasized generating modifiers that resonate with the theme of the initial prompt. This strategy enables the model to produce distinct innovations tailored to the nature of the prompt.

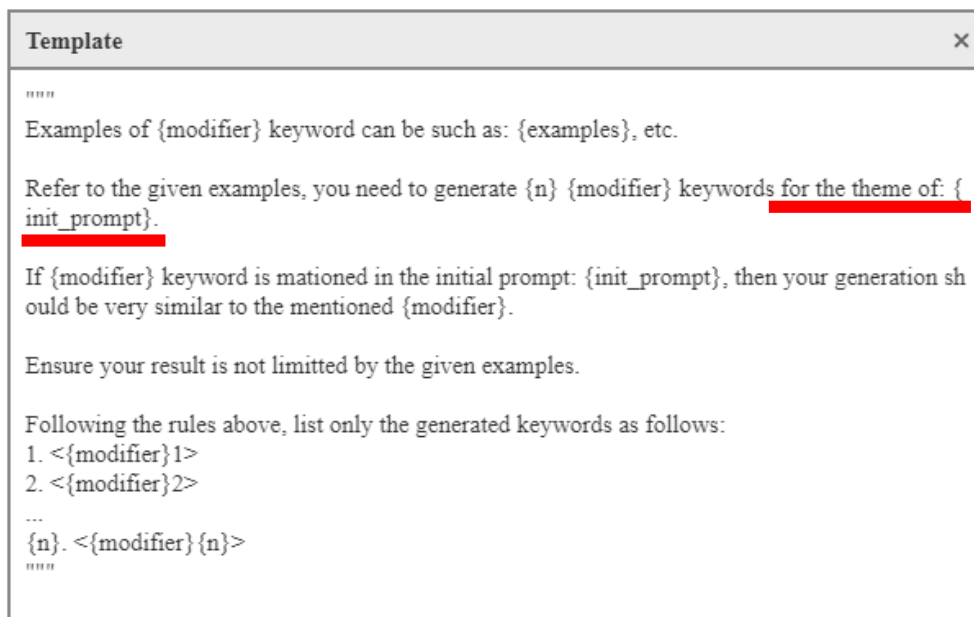


Figure 4.8: Prompt template in prompt engineering

## 4.2. Comparative analysis focusing on the aesthetic differences between images generated with and without the inclusion of prompt modifiers

### 4.2.1. Objective

The primary aim of this experiment was to investigate the aesthetic differences between images generated from initial prompts and those generated from optimized prompts by our prompt optimization system--Genjourney.

### 4.2.2. Setup

We set 20 phrases as initial prompts and used “Midjourney Model V5.2” model for half of the prompts and the “Niji Model V5” for the other half to generate images as shown in Table 4.2.

Midjourney Model V5.2	“A clash of elemental Titans”
	“A day in the life of a robot”
	“A frozen wilderness”
	“A knight facing a dragon on a mountaintop”
	“Fantasy magic school in a hidden valley”
	“Futuristic hero”

Niji Model V5	"Futuristic metropolis with flying cars"
	"Monster hunter"
	"Sci-Fi space station orbiting an alien planet"
	"The final frontier"
	"A strong queen"
	"Eastern dragon"
	"Greek god"
	"Iron man"
	"Mecha"
	"Starbuck cup design"
	"Transformer"
	"White hair girl"
	"Wizard"
	"World of Xianxia"

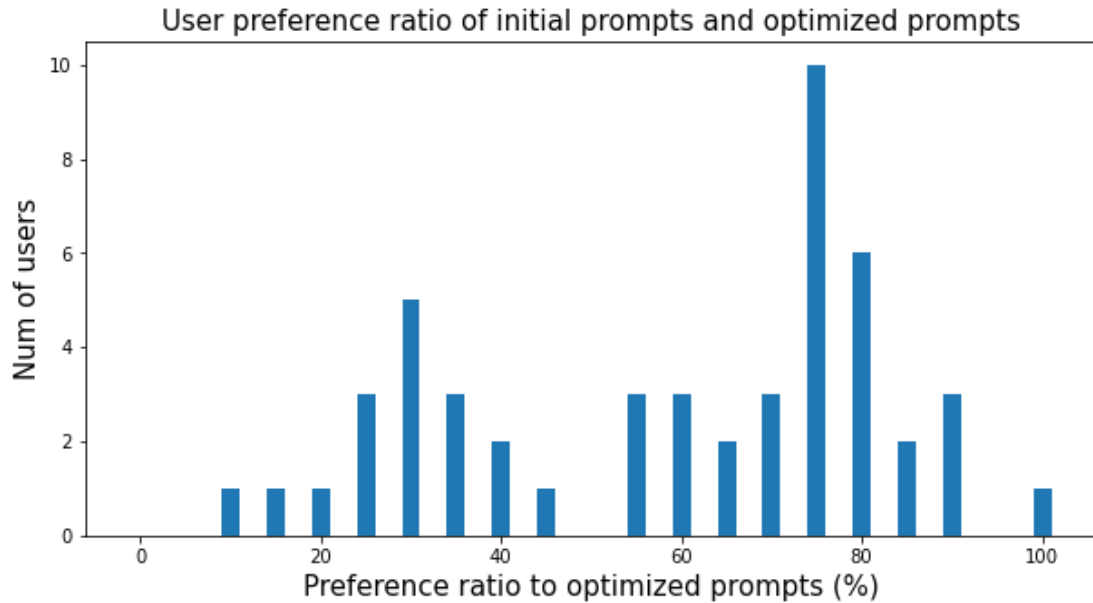
**Table 4.2: 20 initial prompts for Midjourney Model V5.2 & Niji Model V5**

For each initial prompt, we generated five sets of images using initial prompt and another five sets using five distinct optimized prompts produced by the Genjourney System, resulting in a total of 100 image sets as samples.

We devised an assessment survey focusing on 100 sets of images. We recruited 50 participants from a crowdsourcing website to conduct this evaluation. Each participant was required to answer 20 multiple-choice questions, each valued at 5 points, selecting the image they deemed the most aesthetically pleasing. Each question offered 10 image options associated with a specific initial prompt, and these options were presented in a randomized order. In this manner, based on the scores obtained, we could precisely gauge user preferences between the initial prompts and the optimized prompts.

### 4.2.3. Results

Survey results in Figure 4.9 showed that the average preference ratio for images from optimized prompts is 59.1%. Moreover, 66% of participants preferred images produced by the prompt optimization system over those generated from initial prompts. Among this group, about 66.7% had a preference level of 75% or higher for the images from optimized prompts.



**Figure 4.9: Preference ratio to images from optimized prompts compared with those from initial prompts**

#### 4.2.4. Analysis

The results with 59.1% preference ratio for images from optimized prompts, as reflected in Figure 4.9, speaks volumes about the potential of the system. This may suggest that while the initial prompt-based images did possess aesthetic qualities, they might have adhered closely to a standardized model-defined style which may not align with the aesthetic preferences of all individuals.

One of the striking differences noted between the two sets of images was the breadth of stylistic diversity. The optimized prompts, evidently, facilitated the generation of images that spanned a wider stylistic spectrum. As illustrated in Table 4.3, while images generated from the initial prompt “white hair girl” maintained aesthetic consistency, they lacked significant variation. Conversely, those crafted from Genjourney-optimized prompts flaunted pronounced differences, particularly in terms of artistic style, color, and lighting nuances.

This preference trend might be attributed to the inherent human inclination towards diversity and novelty. Moreover, while this experiment doesn't delve deeply into it, individual participant backgrounds, personal experiences, and cultural influences might have also played a role in their selections. Such dimensions warrant a deeper exploration in subsequent studies.

---

Images Generated by Initial Prompt

---

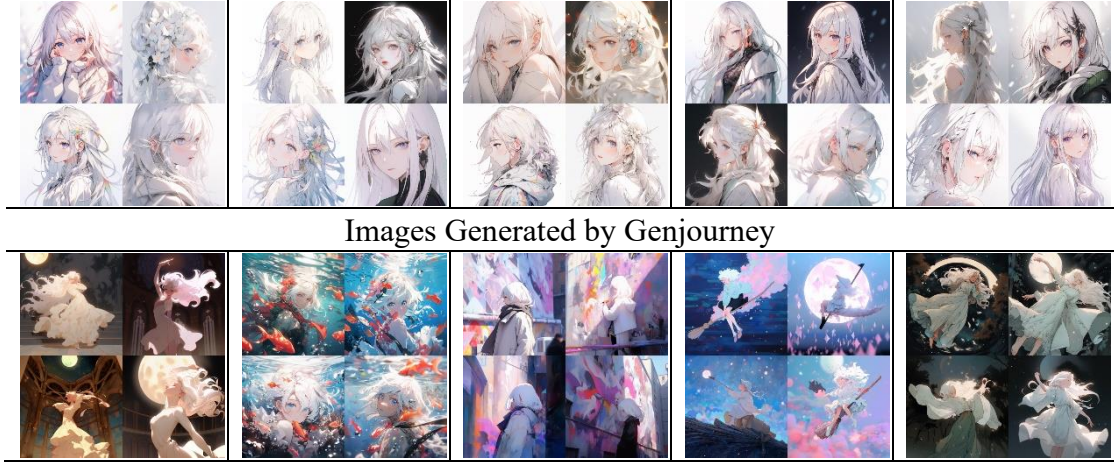


Table 4.3: Images instance generated by initial prompt and Genjourney

### 4.3. Assessment of the prompt optimization system's capacity to evolve towards a desired target image

#### 4.3.1. Objective

The primary objective of this experiment was to assess the capability of our prompt optimization system in evolving towards a desired target image, based on users' specific expectations which are often rooted in their descriptions or pre-existing image samples.

#### 4.3.2. Setup

##### 4.3.2.1. Target images and initial prompts setting

- Chose 10 representative target images each from “Midjourney Model V5.2” and “Niji Model V5”, resulting in a total of 20 images from Prompt Hero [36].
- Assigned a brief description to each image as initial prompt, limited to three phrases.

##### 4.3.2.2. Quantitative comparative approach

- Utilized the describe function from Midjourney to inversely convert target images into textual prompts.
- Employed the pre-trained “all-mpnet-base-v2” model, based on the MPNet pre-training method, to compute the semantic similarity between the initial prompts and the inversely generated target prompts [13, 40].
- By improving the semantic similarity, we hope to make the generated image closer to the visual and aesthetic effects of the target image.

##### 4.3.2.3. Optimization and Similarity Assessment

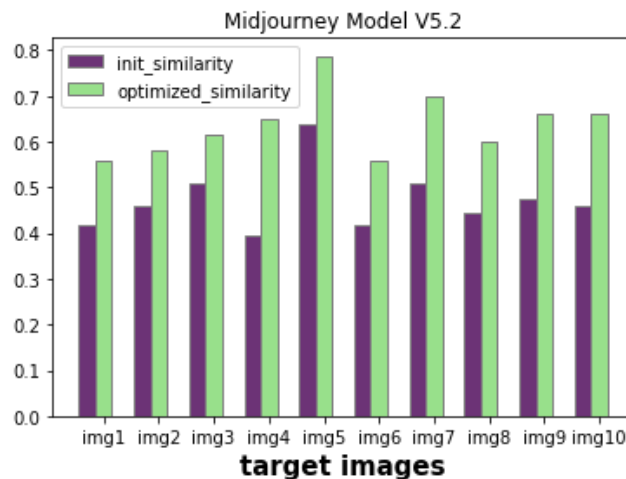
- Compared the semantic similarity between initial and target prompts.
- Optimized the initial prompts using Genjourney with mutation rate set as “medium”, generating five prompts each time, and set the semantic similarity as

average user feedback score. For example, if semantic similarity is 88%, then average feedback score represented by semantic similarity is 8.8 out of 10.

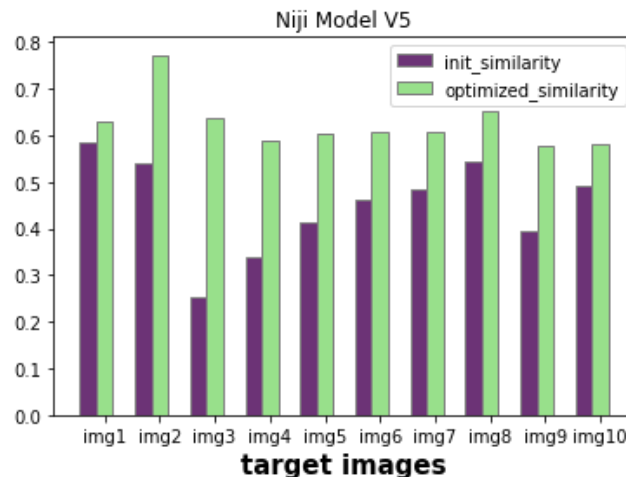
- Assessed the semantic similarity of these prompts in 10 generations with the target prompts to find the prompt with maximum semantic similarity.
- Repeated the above steps 5 times and calculated the average maximum semantic similarity.

### 4.3.3. Results

The chart of Figure 4.10 and Figure 4.11 clearly indicates that the prompt optimization system effectively optimizes the initial prompts in semantic similarity with target images. For the "Midjourney Model V5.2", there was a 16% improvement in semantic similarity, from 47% to 64%. For the "Niji Model V5", the similarity increased by approximately 17%, from 45% to 62%. A visual representation in Table 4.4 will further showcase images generated from the initial prompt, the optimized prompt with the highest semantic similarity and the target image.



**Figure 4.10: Semantic similarity of initial prompts and most optimized prompts with target prompts for Midjourney Model V5.2**





**Figure 4.11: Semantic similarity of initial prompts and most optimized prompts with target prompts for Niji Model V5**

Index	Image from Initial Prompt	Image with Highest Semantic Similarity	Target Image
1			
<p>Initial prompt: “lace pattern”</p> <p>Optimized prompt: “lace pattern, enchanting fairytale, magical forest, whimsical creatures, mystical adventure, childhood nostalgia, Abstract expressionism, René Lalique, Dusty rose, Close-up detail shot of intricate lace pattern, Volumetric lighting on lace pattern, Lace pattern rendered with Arnold renderer”</p>			
2			
<p>Initial prompt: “stylish building design”</p> <p>Optimized prompt: “stylish building design, colonial-inspired, white facades, wooden shutters, central courtyard, lush tropical gardens, Neo-Futuristic, Bjarke Ingels, Ivory cream, Asymmetrical architectural design, Eclectic color washes, Material texture rendering”</p>			
3			
<p>Initial prompt: “Zeus, golden armor”</p> <p>Optimized prompt: “Zeus, golden armor, <b>lightning storms</b>, divine wrath, heavens trembling, Renaissance, Ken Kelly, Apollo Orange, Panoramic view of Zeus wielding his golden armor, Glittering armor sheen, Physically accurate rendering”</p>			



4



Initial prompt: “a woman wearing golden headgear, cyan color”

Optimized prompt: “a woman wearing golden headgear, cyan color, mounted on a magnificent winged horse, **Surrealism**, Jozef Israels, Golden aquamarine, three views of woman with cyan-colored headgear, Golden rim lighting, **3DCG modeling**”

5



Initial prompt: “a blueberry cake”

Optimized prompt: “a blueberry cake, irresistible aroma, golden and crispy crust, garnished with mint leaves, presented on a modern dessert platter, Contemporary realism, Roy Lichtenstein, Sapphire sparkle, Macro lens shot of blueberry cake, Blueberry crepuscular ray, Global illumination”

6



Initial prompt: “a young woman, wearing futuristic bionic clothes, crab claws on head”

Optimized prompt: “a young woman, wearing futuristic bionic clothes, crab claws on head, interacting with a floating robot companion, Abstract expressionism, H.R. Ford, Laser lemon, Detail shot of young woman's bionic clothes and crab claws on head, Volumetric lighting on crab claws, Real-time rendering”

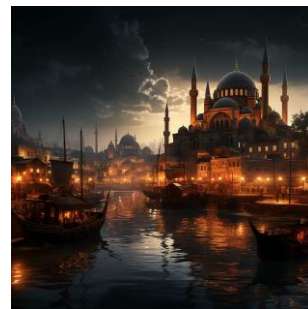
7



Initial prompt: “An astronaut, exploring among meteorites”

Optimized prompt: “An astronaut, exploring among meteorites, constructing a temporary base on an asteroid, Art nouveau, Chiara Bautista, Moon dust beige, Fisheye lens view of astronaut in space, Lunar eclipse shadow, Real-time rendering”

8



Initial prompt: “Catholic church port, illuminated”

Optimized prompt: “Catholic church port, illuminated, stained glass windows, **evening sky**, tranquil waters, Byzantine, Hieronymus Bosch, Radiant orange, Religious icon detail, Ethereal glow, Photorealistic rendering”

9



Initial prompt: “a white-hair man, with sunglasses and suit, walking in flowers”

Optimized prompt: “a stylish white-haired man, wearing trendy sunglasses and a stylish suit, gracefully making his way through a mesmerizing display of flowers in full bloom, Geometric abstraction, Edvard Munch, Bouquet burgundy, Wide-angle shot of white-hair man with sunglasses and suit in a flower field, Flowery backlight, Realistic shadows”



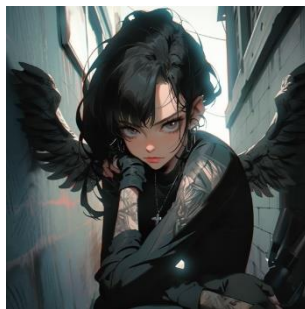
10



Initial prompt: “magic rainforest”

Optimized prompt: “magic rainforest, **luminescent mushrooms**, ethereal music, sparkling fireflies, mystical fog, enchanted creatures, Enigmatic plant life, Shaun Tan, Emerald Green, Product view, Foggy illumination, Motion blur”

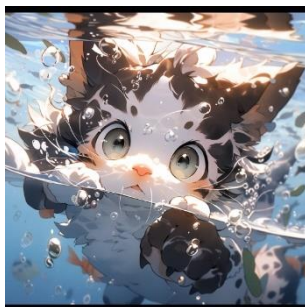
11



Initial prompt: “a black hair cool girl with tattoo”

Optimized prompt: “a rebellious girl with black hair and angel wings tattoos, spreading her message of freedom through street art, Pop art, Takashi Murakami, Midnight black, Fisheye lens shot of black hair cool girl with tattoo, Intense rim lighting accentuating the black hair and tattoos, Realistic hair shading”

12



Initial prompt: “a cat underwater”

Optimized prompt: “a cat underwater, gracefully swimming, surrounded by playful dolphins, diving into an underwater canyon, fascinated by the hidden underwater world, Op art, Jhon Berkey, Abyssal black, Simulated camera perspective capturing a cat underwater, Submerged intense backlight, **Photorealistic textures**”

13



Initial prompt: “a giant tree”

Optimized prompt: “a giant tree, with twisted vines, jungle atmosphere, exotic plants, hidden paradise, vibrant ecosystem, Anime tree, Thomas Moran, Mossy brown, foliage close-up, Tree silhouette against stormy sky, Ray Tracing”

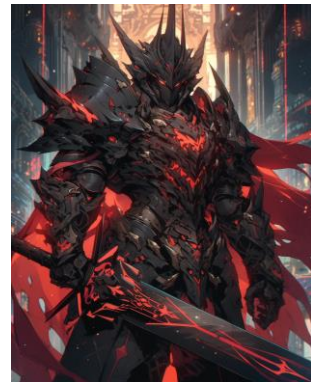
14



Initial prompt: “a future soldier, with helmet and armor”

Optimized prompt: “a futuristic soldier, sporting a **high-tech** helmet and resilient armor, standing tall amidst the ruins of a conquered city, Digital, Simon Gane, Metalloid silver, Panoramic view of soldier's face with helmet, **Cyberpunk** glow in the dark, Detailed armor reflections”

15



Initial prompt: “a knight in black and red armor, holding a sword”

Optimized prompt: “a knight in black and red armor, holding a sword, battling a fearsome dragon, Symbolism, Yoji Shinkawa, Shadow black, Extreme close-up of knight's sword and the intricate design on it, Soft red lighting creating an **ominous** atmosphere around the knight, Translucency”



16



Initial prompt: “a hooded white-hair girl, with face mask”

Optimized prompt: “a hooded white-haired girl, with face mask, painting vibrant graffiti on a city wall, Digital art, Yoshitaka Amano, Icy lavender, Fisheye lens shot of hooded white-hair girl with face mask., Front lighting highlighting the features of the hooded girl, High-definition textures”

17



Initial prompt: “a Chinese swordswoman”

Optimized prompt: “tattoo design of two dragons, fierce, fighting atop a stormy sea, with crashing waves and lightning strikes, Japanese woodblock print, Studio Ghibli, Cobalt blue, Detailed shot of dragon tattoos, Dragon breath illumination, Tattoo shading techniques”

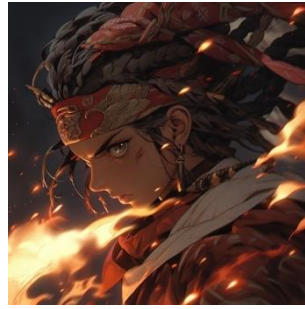
18



Initial prompt: “tattoo design of two dragons”

Optimized prompt: “a cyberpunk-inspired female robot, her glowing synthetic skin reflecting the iridescent wings of the butterflies that surround her, Psychedelic cyberpunk, Takayuki Takeya, Mechanical turquoise, Abstract composition of cyberpunk female robot and butterflies, Cybernetic lens flare, Photorealistic rendering”

19



Initial prompt: “a brave warrior, wearing a headscarf”

Optimized prompt: “a brave warrior, wearing a headscarf, defending a village from a dragon attack, flames soaring high, bravery in every step, Art Deco, Takehiko Inoue, Headdress ebony, Dramatic **close-up** on a warrior's face, Studio-lit warrior, detailed warrior model”

20



Initial prompt: “a colorful anime building”

Optimized prompt: “a colorful anime building, serene riverside, elegant swans, blooming water lilies, peaceful boat rides, picturesque sunset views, anime-inspired Renaissance, Chiara Bautista, Sky blue, Faceshot of a colorful anime building, Anime Volumetric Lighting, Non-photorealistic rendering”

**Table 4.4 Images generated from the initial prompt, the optimized prompt with the highest semantic similarity and the target image**

#### 4.3.4. Analysis

The optimization system often faces challenges due to the inherent subjectivity embedded within initial prompts [44]. It's vital for the system to detect and respond to subtle or hidden user preferences effectively. In this context, the prompt optimization system's ability in approximating the target image, even when the initial prompts were subjective and the approximation was not obvious in some cases, is noteworthy.

For instance, the fourth target image, which has surrealistic features, was enhanced post-optimization with the term "surrealism". Similarly, for the fifteenth and seventeenth target images, Genjourney introduced elements like "luminescent mushrooms" and "dragon" which are elements of the target images, even though they weren't explicitly mentioned in the initial prompts, due to the effect of prompt template in Figure 4.8. Some other Bold and red text in the chart also demonstrates this ability of Genjourney.

The results imply that even with minimal input, users can harness the power of

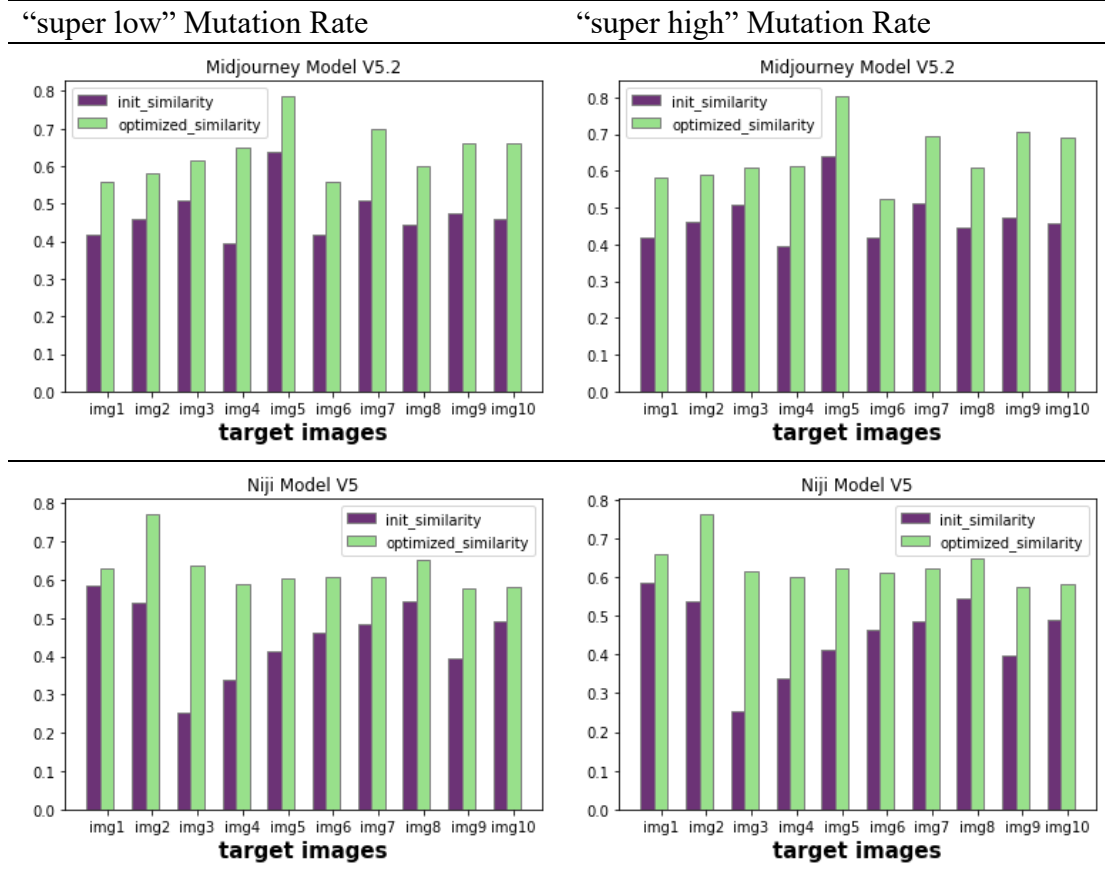
Genjourney to drive innovation. Genjourney 's ability to introduce themes or styles that align with the essence of the initial prompt, even if not explicitly mentioned, is a testament to its potential. This capability can empower users to achieve distinct aesthetic and visual outcomes, even if they provide only a basic subject or theme.

For users, this means a more intuitive and user-friendly experience. They don't need to be overly specific or detailed in their prompts. The system can fill in the gaps, making the image generation process more seamless and aligned with their vision.

#### 4.3.5. Expansion Experiment of Mutation Rate

In previous experiments, we set the mutation rate of Genjourney to 'medium' with mutation coefficient as 0.3. In this section, we will further investigate the impact of varying this rate on the semantic similarity observed in our study.

After adjusting the variation rate to both 'super low' and 'super high' with mutation coefficient as 0.1 and 0.6, setting number of generations as 3 and replicating the procedures outlined in section 4.3.2, we obtained the semantic similarity results as depicted in the Table 4.5 and 4.6.



**Table 4.5: Semantic similarity of initial prompts and most optimized prompts with target prompts for Midjourney Model V5.2 and Niji Model V5 with different mutation rates.**

	Midjourney Model V5.2	Niji Model V5
“super low” Mutation Rate	16.4%	17.5%
“super high” Mutation Rate	16.9%	17.9%

**Table 4.6: Increase semantic similarity between initial prompts and most optimized prompts with target prompts for Midjourney Model V5.2 and Niji Model V5 with three different mutation rates.**

Data from Table 4.6 indicates that even for mutation rates designated as "super high" and "super low," the difference of increase semantic similarity for these two mutation rates is still small (less than 1%) when the number of generations is small. Furthermore, the semantic similarity difference between these rates and the "medium" mutation rate remains within 1%. This suggests a ceiling effect on the improvement of semantic similarity. A close analysis of Figure 4.10 and Table 4.5 reveals a significant correlation between post-Genjourney optimization semantic similarity and its initial values. For every initial prompt, the enhancement in semantic similarity is typically constrained. The final semantic similarity largely depends on the likeness between the initial and target prompts.

## **4.4. User Satisfaction Survey for Prompt Optimizing System**

### **4.4.1. Objective**

This section aims to evaluate the effectiveness of Genjourney in enhancing image generation based on user preferences. With the proliferation of AI-driven image generation tools, ensuring that generated images align closely with user expectations has become paramount.

### **4.4.2. Setup**

#### **4.4.2.1. Survey Design and Collection**

- Surveyed 50 participants via a crowdsourcing platform.
- Collected data on initial prompts, artistic scores of each generation, averaged across five dimensions: subject, artistic style, color palette, compositional perspective, and lighting effect, which is the average value of user feedback score and the image that best aligned with the user's aesthetic preference.

#### **4.4.2.2. Image Generation and Rating**

- Presented users with six initial prompts.
- Generated ten generations of images for half using the “Midjourney Model V5.2” and the other half using “Niji Model V5.”
- Participants rated each generation of images, resulting in 300 data samples.



### 4.4.3. Results

Figure 4.10 and Figure 4.11 show a clear upward trend in average scores, with an increase of approximately 1.15 points from the 0th to the 10th generation. While individual trends varied based on the initial prompts, the general trend was positive. The data also indicated that in 95% of instances, subsequent generations had higher scores than the initial generation, with the 10th generation outperforming the initial in 70% of cases.

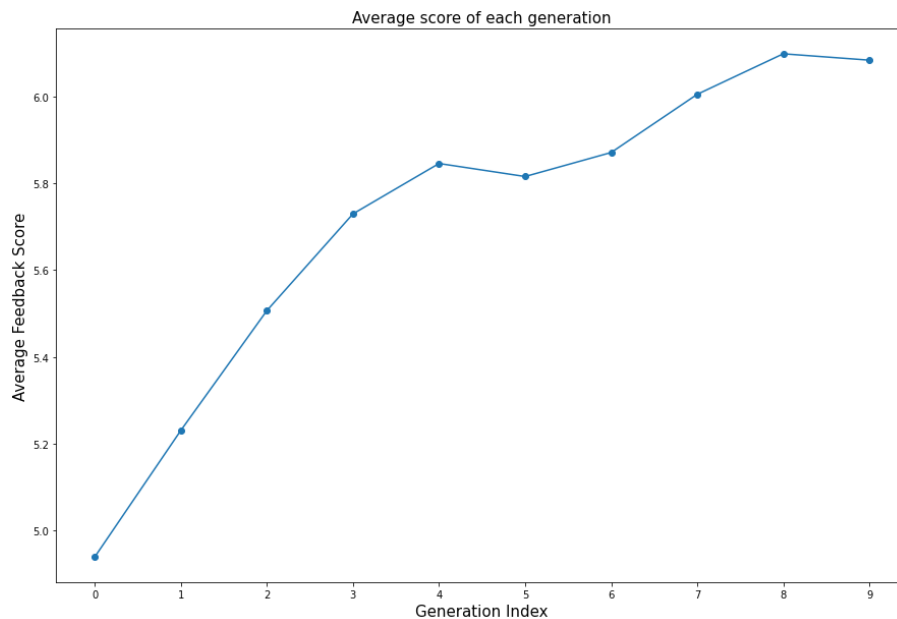


Figure 4.10: Overall average artistic score for each generation

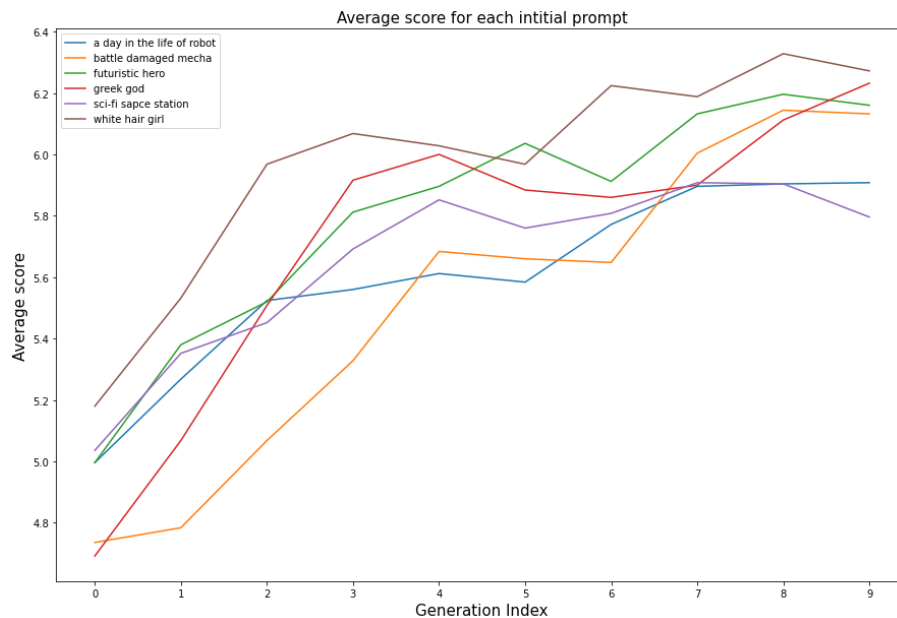


Figure 4.11: Average artistic score for each generation of six initial prompts

### 4.4.4. Analysis

The upward trend in average scores from the initial to the 10th generation suggests that Genjourney is not only effective in refining images but does nearly consistently across multiple iterations. This consistency is indicative of a robust optimization algorithm that learns and adapts effectively from user feedback. While the general trend was positive, there were instances where certain prompts or image generations deviated from the expected results. Delving into these outliers can offer insights into areas where the system may face challenges.

The fact that in 95% of instances, subsequent generations outscored the initial generation is a testament to the system's iterative refinement capability. The 10th generation's performance, in particular, highlights the system's potential in converging towards a user's aesthetic ideal over time.

The inclusion of multiple artistic dimensions (subject, style, color palette, etc.) in the scoring mechanism provides a granular view of user preferences. The system's ability to enhance images across these dimensions suggests a holistic approach to optimization, catering to diverse aesthetic sensibilities.

These results have broader implications for real-world applications of the prompt optimization system. For instance, in industries like advertising or content creation, where aligning visual content with audience preferences is crucial, such a system could be invaluable. In advertising, for instance, tailoring an ad's visual elements to a specific demographic could lead to higher engagement rates. Using a system like Genjourney, advertisers could refine their visuals based on preliminary audience feedback, ensuring a more impactful campaign launch."

## **5. Conclusion**

### **5.1. Summary**

The rapid advancement of AI image generation tools has underscored the need for systems that can optimize textual prompts, ensuring that generated images align more closely with user preferences. In response to this need, our research introduces Genjourney, a semi-automated system designed to optimize textual prompts for AI image generation models. This system facilitates user interaction with the generated images and employs an interactive evolutionary computation approach, specifically a genetic algorithm, to refine prompts and images in a direction that better aligns with user aesthetics.

Through a series of meticulously designed experiments, we evaluated the system's capacity to generate a diverse array of prompt modifiers, its impact on the aesthetic quality of the images produced, its ability to evolve towards a target image, and overall user satisfaction. Our findings indicate that Genjourney can achieve a high uniqueness rate when generating prompt modifiers, both for the same initial prompt and for

different ones, showcasing its diversity due to prompt engineering. Furthermore, images generated from optimized prompts produced by Genjourney are more favored by users compared to those generated from initial prompts.

Additionally, the Genjourney system demonstrated its capability to automatically approximate a target image based on semantic similarity. Even with succinct initial prompts, Genjourney adeptly introduces aesthetic elements that align with the prompt's theme, such as subject and style which are not-mentioned elements in the target images. This capability not only empowers users to achieve distinct aesthetic and visual outcomes but also underscores the system's creative and expansive potential with minimal initial input.

Lastly, we conducted a user satisfaction survey regarding the system. Our results revealed an overall upward trend in the aesthetic scores users assigned to images as the number of generations in Genjourney increased. Compared to the initial generation, users often found images from subsequent generations to be more in line with their aesthetic preferences, further highlighting the system's efficacy in refining images through optimized prompts.

## **5.2. Limitation**

While the Genjourney system has showcased promising results, it is not without its limitations.

### **5.2.1. Limitation on Semantic Similarity Improvement**

Although Genjourney can automate the enhancement of semantic similarity between optimized prompts and target image prompts, the improvement in semantic similarity from the fifth to the tenth generation is within a margin of 1%. Moreover, adjusting the mutation rate of prompt modifiers has minimal impact on the results, indicating that there's a ceiling to how much semantic similarity can be improved.

### **5.2.2. Subjectivity of Aesthetic Preferences**

The inherent subjectivity of aesthetic preferences means that while the system can evolve prompts based on diverse user aesthetics, it might not always align perfectly with individual tastes.

### **5.2.3. Model Constraints**

The GPT-3.5-turbo model's training data is limited to information up to 2021, making it unaware of events or developments post-2021. Additionally, the Midjourney model's prompt data is not within the knowledge base of the GPT-3.5-turbo model. Given that Midjourney is a proprietary model, we lack insights into its specific language training

data. Consequently, our prompt structure is derived from the official Midjourney documentation, AI image generation researchers, and community content, rather than being based on Midjourney's language training data.

#### **5.2.4. Dependence on Initial Prompts**

The performance of the Genjourney system is, to some extent, contingent on the quality and clarity of the initial prompts. Ambiguous, overly abstract, or broad prompts might necessitate more generations to yield optimal results.

### **5.3. Recommendation**

#### **5.3.1. Utilizing More Advanced Large Language Models for Prompt Modifier Generation**

In this study, we employed the GPT-3.5-turbo model as our primary large language model for generating prompt modifiers. Future research could consider leveraging more advanced models, such as GPT-4 or CLAUDE-2, for this purpose. These models, equipped with larger training datasets, more parameters, and updated knowledge bases, might offer superior performance in language generation [9, 31].

#### **5.3.2. Fine-tuning Models for User Personalization**

Fine-tuning is a pivotal aspect of prompt engineering. It enables the model to enhance its performance for specific tasks based on provided samples, building upon its original training [50]. Capitalizing on this feature, we can use the best image prompts and best images generated by users as sample data. By fine-tuning both the large language model and the image generation model with this data, the system can more effectively learn and adapt to individual aesthetic preferences, paving the way for a more personalized user experience [8].

## **References**

- [1] Alam, T., Qamar, S., Dixit, A. and Benaida, M. (2020). Genetic Algorithm: Reviews, Implementations, and Applications. arXiv preprint arXiv:2007.12673. [Online]. Available: <https://arxiv.org/ftp/arxiv/papers/2007/2007.12673.pdf> .
- [2] AlgoWriting. (2020). A simple guide to setting the GPT-3 temperature. Medium. [Online]. Available: <https://algowriting.medium.com/gpt-3-temperature-setting-101-41200ff0d0be> .
- [3] Bentley, P.J. and Corne, D.W. (2002). An introduction to creative evolutionary systems. In: Bentley, P.J. and Corne, D.W. (eds.) Creative Evolutionary Systems.

Morgan Kaufmann.

- [4] Bing CHENG. Artificial Intelligence Generative Content (AIGC) Including ChatGPT Opens a New Big Paradigm Space of Economics and Social Science Research[J]. China Journal of Econometrics, 2023, 3(3): 589-614 <https://doi.org/10.12012/CJoE2023-0032> .
- [5] Boehman, C. (2023). Midjourney vs. DALL-E vs. Stable Diffusion: Which Is Better? MakeUseOf. [Online]. Available at: <https://www.makeuseof.com/midjourney-vs-dalle-vs-stable-diffusion/> .
- [6] Branwen, G. (2020). GPT-3 Creative Fiction. Gwern.net. [Online]. Available at: <https://gwern.net/gpt-3> .
- [7] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S. and Sun, L. (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. ArXiv. [Online]. Available: <https://arxiv.org/abs/2303.04226> .
- [8] Chen, W. et al. (2023). Subject-driven text-to-image generation via apprenticeship learning. arXiv preprint arXiv:2304.00186. [Online]. Available at: <https://arxiv.org/pdf/2304.00186.pdf> .
- [9] Claude-2. [Online]. Available at: <https://www.anthropic.com/index/claude-2> .
- [10] David, J. (2023). How to write a good prompt for Midjourney. Abyssale. [Online]. Available: <https://www.abyssale.com/creative-automation/how-to-write-a-good-prompt-for-midjourney> .
- [11] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems.
- [12] Holland, J.H. (1992). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The MIT Press. [Online]. Available: <https://doi.org/10.7551/mitpress/1090.001.0001> .
- [13] Hugging Face. (2021). all-mpnet-base-v2 model. [Online]. Available at: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2> .
- [14] Kingma, D.P. and Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114. [Online]. Available:

<https://arxiv.org/pdf/1312.6114.pdf> .

- [15] Khader, F., Müller-Franzes, G., Tayebi Arasteh, S. et al. (2023). Denoising diffusion probabilistic models for 3D medical image generation. Scientific Reports, 13, 7303. [Online]. Available: <https://doi.org/10.1038/s41598-023-34341-2> .
- [16] Ko, H.-K., Park, G., Jeon, H., Jo, J., Kim, J. and Seo, J. (2022). Large-scale Text-to-Image Generation Models for Visual Artists' Creative Works. arXiv preprint arXiv:2210.08477. [Online]. Available: <https://arxiv.org/pdf/2210.08477.pdf> .
- [17] Kosorukoff, A. (2001). Human based genetic algorithm. In: Human-based Genetic Algorithm, pp. 3464–3469. ISBN 978-0-7803-7087-6. doi:10.1109/ICSMC.2001.972056 .
- [18] Li, C. (2020). OpenAI's GPT-3 Language Model: A Technical Overview. [Online]. Lambda Labs. Available at: <https://lambdalabs.com/blog/demystifying-gpt-3> .
- [19] Midjourney. (2023). [Online]. Available at: <https://www.midjourney.com/app/> .
- [20] Midjourney. (2023). Blend. [Online]. Available at: <https://docs.midjourney.com/docs/blend> .
- [21] Midjourney. (2023). Midjourney Documentation and User Guide. [Online]. Available: <https://docs.midjourney.com/docs/prompts> .
- [22] Midjourney. (2023). Midjourney Showcase. [Online]. Available: <https://www.midjourney.com/showcase/recent/> .
- [23] Midjourney. (2023). Prompts. [Online]. Available: <https://docs.midjourney.com/docs/prompts> .
- [24] Midjourney. (2023). Version. [Online]. Available at: <https://docs.midjourney.com/docs/model-versions> .
- [25] Mittal, A. (2023). The Essential Guide to Prompt Engineering in ChatGPT. Unite AI. [Online]. Available: <https://www.unite.ai/prompt-engineering-in-chatgpt/> .
- [26] Nakanishi, Y. (1996). Applying evolutionary systems to design aid system. In: Proceedings of Artificial Life V (Poster Presentation), pp. 147–154.
- [27] O'Connor, R. (2022). How DALL-E 2 Actually Works. AssemblyAI. Available at: <https://www.assemblyai.com/blog/how-dall-e-2-actually-works/> .
- [28] OpenAI. (2021). CLIP: Connecting Vision and Language with Localized

- Narratives. arXiv preprint arXiv:2103.00020. [Online]. Available: <https://arxiv.org/abs/2103.00020v1>.
- [29] OpenAI. (2023). API Reference. [Online]. Available at: <https://platform.openai.com/docs/api-reference>.
- [30] OpenAI. (2023). CLIP: Connecting text and images. OpenAI. [Online]. Available: <https://openai.com/research/clip>.
- [31] OpenAI. (2023). GPT-4 Technical Report. [Online]. Available at: <https://cdn.openai.com/papers/gpt-4.pdf>.
- [32] Oppenlaender, J. (2022). A Taxonomy of Prompt Modifiers for Text-To-Image Generation. arXiv preprint arXiv:2204.13988. [Online]. Available: <https://arxiv.org/abs/2204.13988>.
- [33] Oppenlaender, J. et al. (2023). Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering.
- [34] Pati, A. and Lerch, A. (2021). Attribute-based regularization of latent spaces for variational auto-encoders. *Neural Computing & Applications*, 33, 4429–4444. [Online]. Available: <https://doi.org/10.1007/s00521-020-05270-2>.
- [35] Pavlichenko, N. and Ustalov, D. (2023). Best Prompts for Text-to-Image Models and How to Find Them. Toloka, Belgrade, Serbia. Available at: <https://arxiv.org/pdf/2209.11711.pdf>.
- [36] Prompt Hero. (2023). Prompt Hero. [Online]. Available: <https://prompthero.com/>.
- [37] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125v1. [Online]. Available: <https://arxiv.org/abs/2204.06125v1>.
- [38] Reed, S., Akata, Z., Lee, H. and Schiele, B. (2016). [Title of the Paper]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 49-58.
- [39] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *CVPR2022*, 10684–10695. [Online]. Available: <https://arxiv.org/pdf/2112.10752.pdf>.
- [40] Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.-Y. (2020). MPNet: masked and permuted pre-training for language understanding. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 16857-

16867.

- [41] Soni, N. (2023). Midjourney: Bridging LLM and Diffusion Models for Smarter Images. LinkedIn. [Online]. Available at: <https://www.linkedin.com/pulse/midjourney-bridging-llm-diffusion-models-smarter-images-naman-soni> Wankhede, C. (2023). What is Midjourney AI and how does it work? Android Authority. [Online]. Available at: <https://www.androidauthority.com/what-is-midjourney-3324590/> .
- [42] Stoimenova, N. and Price, R.A. (2020). Exploring the Nuances of Designing (with/for) Artificial Intelligence. *Design Issues: history/theory/criticism*, 36(4), 45-55. [Online]. Available: [https://doi.org/10.1162/desi\\_a\\_00613](https://doi.org/10.1162/desi_a_00613) .
- [43] Streamlit. (2023). [Online]. Available at: <https://streamlit.io/> .
- [44] Takagi, H. (1998). Interactive Evolutionary Computation: System Optimization Based on Human Subjective Evaluation. In: *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, 1998, pp. 17-19.
- [45] The Next Leg. (2023). [Online]. Available at: <https://www.thenextleg.io/> .
- [46] Thompson, A.D. (2022). GPT-3.5 + ChatGPT: An illustrated overview. [Online]. Life Architect. Available at: <https://lifearchitect.ai/chatgpt/> .
- [47] Vartiainen, H. and Tedre, M. (2023). Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity*, 34(1), pp. 1-21.
- [48] Vass, K. (2023). HAROLD COHEN: ‘ONCE UPON A TIME THERE WAS AN ENTITY NAMED AARON’. Kate Vass Galerie. [Online]. Available at: <https://www.katevassgalerie.com/blog/harold-cohen-aaron-computer-art> .
- [49] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015). Show and tell: A neural image caption generator. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/CVPR.2015.7298935 .
- [50] Ward, K., Chen, C.Z. and Ma, Y., 2021. Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27(6), pp.763-778.
- [51] Zhang, C., Zhang, C., Zhang, M. and Kweon, I.S. (2023). Text-to-image Diffusion Models in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909v2*. [Online]. Available at: <https://doi.org/10.48550/arXiv.2303.07909> .