Master's Thesis

Text-to-Image Generation through Interactive Prompt Visual Exploration

Bofei Huang

Supervisor Haoran Xie

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Knowledge Science)

March, 2025

Abstract

In recent years, generative Artificial intelligence (AI) has demonstrated significant progress in text-to-image generation, merging advancements in computer vision and natural language processing. As above, cutting-edge models such as Stable Diffusion and DALL·E synthesize high-quality images with stunning aesthetics, offering unprecedented opportunities for creativity and innovation. However, crafting prompts that accurately reflect user intent and align with the model's interpretation remains a significant challenge, particularly for novice users lacking prompt engineering expertise. This challenge is further compounded when users aim to explore their creativity and generate ideas iteratively, as the current workflow often requires repeated modifications and refinements, resulting in a time-intensive trial-and-error process. These limitations highlight the need for tools to bridge the gap between user intent and generative model outputs, enabling experts and non-experts to harness the potential of generative AI for creative tasks effectively.

To address these issues, PromptNavi is proposed as an interactive system designed to assist users in enhancing and refining prompts through visual exploration and iterative optimization within text-to-image generative models. PromptNavi introduces a novel approach to prompt refinement, leveraging an attribute interpolation system powered by large language models. This system analyzes initial user inputs and suggests enhancements to align prompts with desired outcomes. A node is a fundamental unit representing the system's image or prompt component. Each node encapsulates visual attributes and semantic information, allowing users to manipulate and refine elements interactively. Furthermore, PromptNavi provides a dynamic, node-based visual interface that transforms the traditionally repetitive and opaque cycle of prompt engineering into an intuitive and interactive experience. Users can adjust prompt attributes, transfer them to generate other images, and iteratively refine their inputs based on visual feedback. The strength of connections between nodes visually represents how modifications to prompts influence the generated outputs, allowing users to understand better and control the relationship between prompts and images than traditional text-to-image generation.

Attributes refer to an image's key visual and semantic properties, such as color, style, and composition. PromptNavi empowers users to discover and integrate these attributes effectively. The system significantly reduces the cognitive and temporal demands of prompt engineering by enabling real-time visual feedback and intuitive attribute manipulation. Additionally, PromptNavi promotes deeper insights into text-to-image generative models, ensuring accessibility for

novice users while providing advanced customization options—such as fine-tuned attribute weighting, multi-node attribute blending, and hierarchical prompt adjustments—for experienced users seeking greater control over image generation. This approach aims to democratize access to generative AI technologies, enabling a broader audience to fully realize their creative potential.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of the development of generative AI, introduces the research objectives, and briefly describes this study's system design and contributions, laying the foundation for subsequent chapters.

1.1 Background

This section introduces the applications of generative Artificial intelligence (AI) in text-to-image generation, highlights challenges in prompt design, and discusses the need for intuitive tools to reduce user difficulties.

1.1.1 Text-to-Image Models

Generative AI has revolutionized creative industries through its state-of-the-art capabilities in text-to-image generation and other domains such as music composition, 3D asset creation, and beyond. Models like Stable Diffusion [7] and DALL·E-3 [8] exemplify this transformation by producing high-quality and visually appealing outputs from natural language descriptions. These breakthroughs have unlocked widespread applications in fields like fashion design, architectural design, and digital art [9] [10] [11], different applications and types can be found in Fig. 1.1. Even in the field of archaeology, Generative AI has still demonstrated its powerful potential. For example, Xie et al. [12]proposed the DiffOBI system by combining text prompts with object images to produce high-quality images that align with the unique characteristics of oracle bone inscriptions. This technology facilitates the recreation of oracle bone art styles and provides new tools for cultural heritage preservation and artifact restoration. However, creating effective prompts to guide image generation remains a significant challenge [13]. Users usually rely on an iterative trial-and-error process involving the composition of

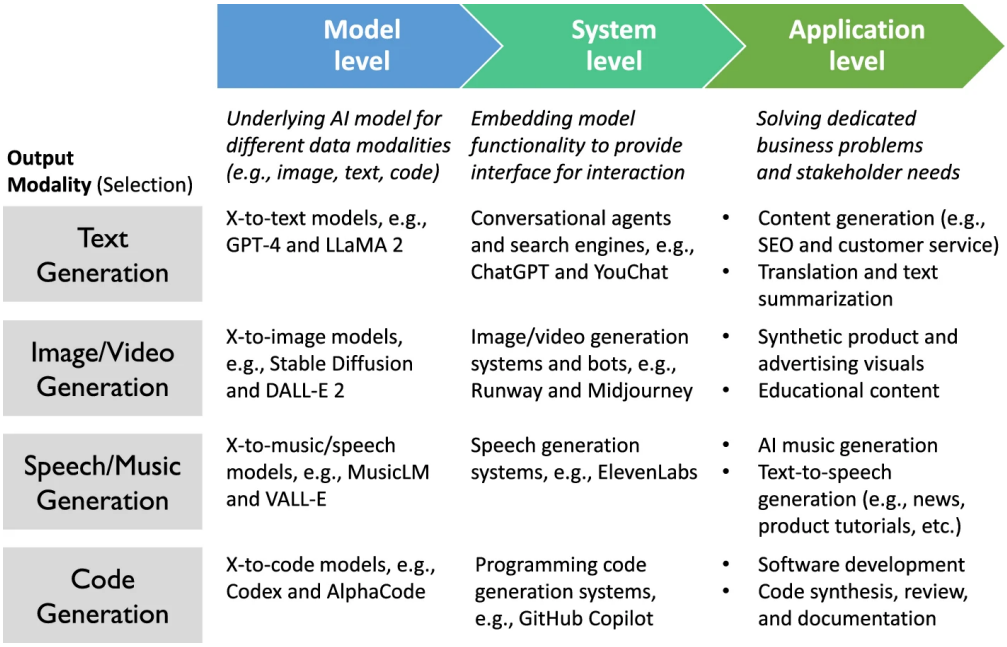


Figure 1.1: A model-, system-, and application-level view on generative AI(adapted from [1])

prompts, analysis of generated results, and refinement of inputs. This workflow is often complex and unintuitive, hindered by users' limited understanding of how prompt structures influence outputs. The predominantly end-to-end nature of image generation complicates the process further, as users must frequently restart from scratch when adjustments fail to yield desired results [14] [15]. Consequently, bridging the gap between users' creative intentions and the capabilities of generative models has emerged as a critical research priority, making prompt engineering an essential aspect of the text-to-image generation workflow [16].

Interactive methods, including alternative prompts, structured templates (e.g., "subject in the style of style name"), and chain-of-thought reasoning, have shown promise in improving prompt design across domains [17] [18] [19]. Retrieval-augmented techniques enhance output consistency and coherence by integrating external knowledge [20]. Yet, significant challenges persist, such as the unpredictable relationship between prompts and generated images and users' difficulties articulating precise creative intentions. These limitations disrupt the creative process and highlight the need for user-centered interaction designs that simplify prompt creation and align with users' goals.

1.1.2 Large Language Model

Large Language Models (LLMs), including Generative Pre-training Transformer(GPT)-based architectures, enable generative AI systems to interpret and respond to natural language prompts effectively. LLMs are essential in text-to-image workflows, serving as a foundation for translating users' textual inputs into meaningful instructions for generative models [18]. They excel in generating structured prompts, adapting to contextual nuances, and assisting in breaking down complex tasks into sequential steps through approaches like chain-of-thought reasoning.

Despite their advanced capabilities, LLMs also face notable limitations in prompt engineering. Users often encounter challenges in understanding the functional scope of LLMs and the impact of specific prompt components on generated outputs [21]. The abstract nature of these models creates a gap between users' intentions and the system's interpretation, making it difficult to identify effective prompt structures. Additionally, LLMs occasionally produce outputs inconsistent with users' expectations, further complicating the creative process and causing inefficiencies.

Recent studies emphasize the importance of user-centered enhancements in LLM interactions to address these issues. Proposed solutions include interactive systems that suggest new keywords, provide alternative phrasing, and leverage retrieval-augmented methods to enhance prompt optimization [17] [20]. These approaches aim to empower users by improving their ability to articulate precise intentions, predict system responses, and refine prompts effectively. These innovations pave the way for more intuitive and productive creative workflows by aligning LLM capabilities with users' needs.

1.1.3 Interaction with Human-Centered-AI

With the gradual development of the generative AI and large language models mentioned earlier, while AI capabilities grow in strength, they also bring worrying aspects. Improper use of AI or flaws in its systems may lead to negative consequences that harm organizations in social, financial, and legal fields [22]. AI is increasingly competing against humans, accompanied by embedded biases (particularly against minority groups), privacy issues, the possibility of AI running out of control, human rights challenges, and more [23] [24].

Therefore, the concept of human-centered AI has been proposed and is gradually becoming an essential topic in the AI field. For example, the Stanford Institute for Human-Centered Artificial Intelligence (HCAI) has proposed three goals for HCAI: "to technologically reflect the depth represented by human intelligence; to enhance, rather than replace, human capabilities; and to focus on the impact of AI on humanity." Xu et al. [2] introduced a preliminary HCAI framework that

includes three main components: "1) ethically compliant design, which creates AI solutions that avoid discrimination, uphold fairness and justice, and do not replace humans; 2) technology that fully reflects human intelligence, further advancing AI to embody the depth of human intelligence; 3) human factors design to ensure that AI solutions are interpretable, understandable, useful, and usable."

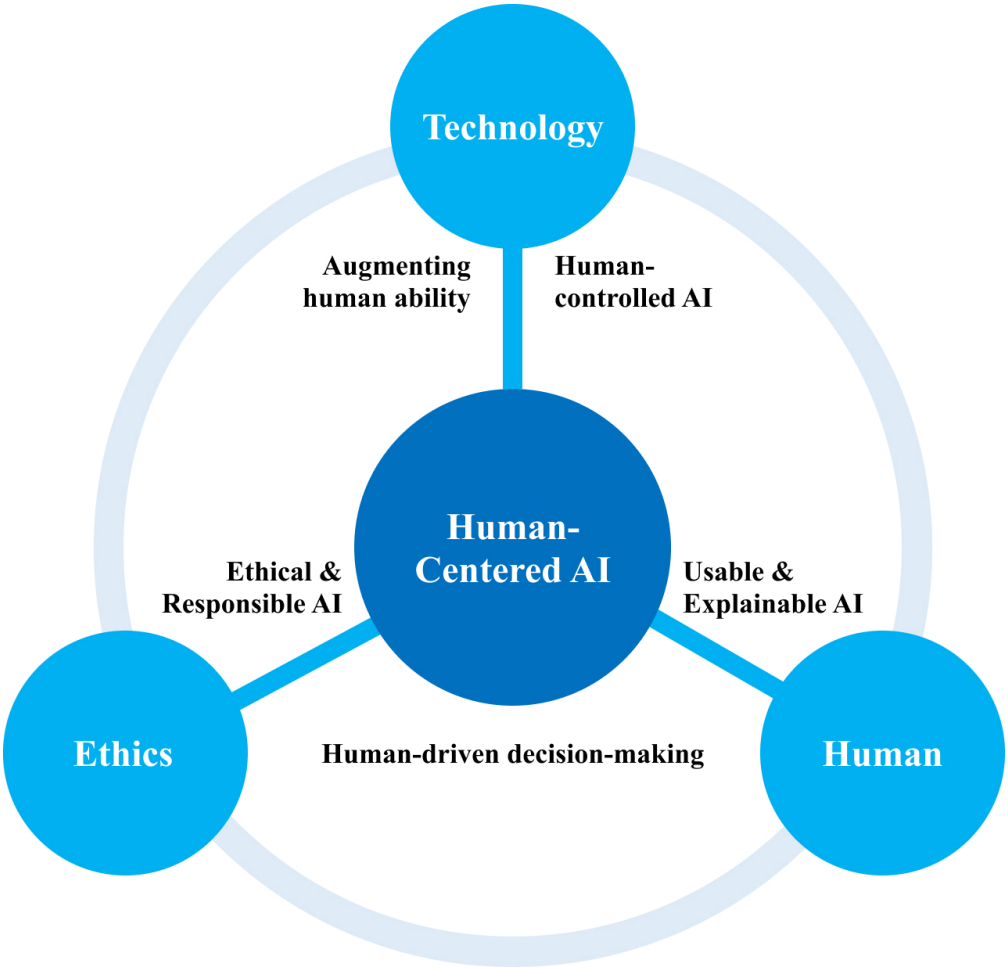


Figure 1.2: The Human-Centered AI (HCAI) framework with specified design goals(adapted from [2])

With the continuous development of HCAI, interaction between humans and AI systems has become a key research focus. This interaction should be based on the complementary capabilities of AI and humans - "AI excels in multitasking, computation, and memory, while humans excel in logical reasoning, language processing, creativity, and emotion." Blackwell [25] emphasizes that while AI systems can mimic human behavior through massive datasets, this "reduces contextual-

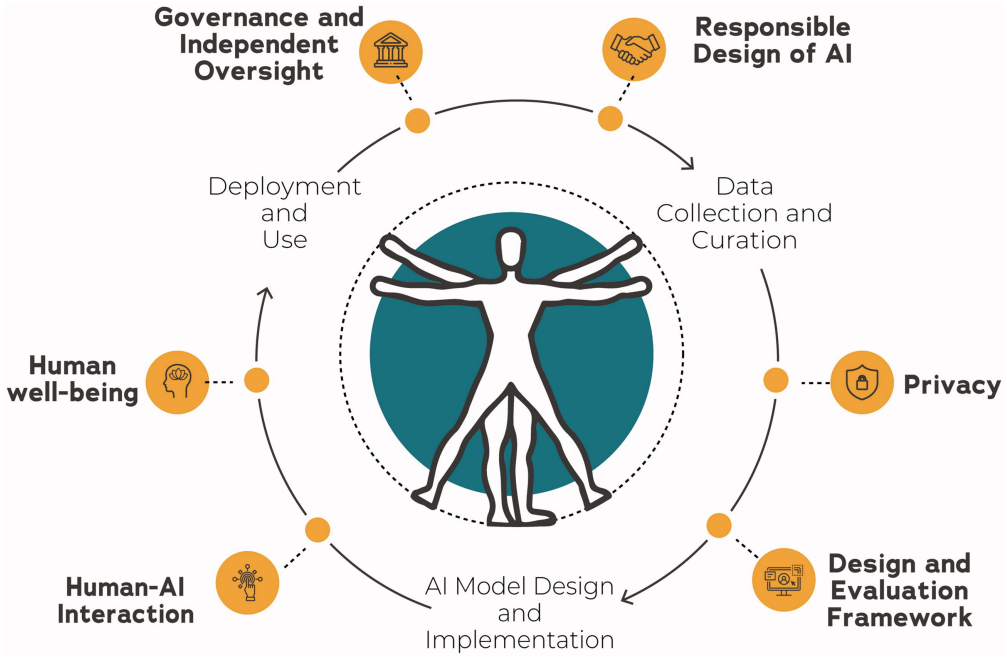


Figure 1.3: HCAI grand challenges.(adapted from [3])

ized humans to machine-like sources of interaction data." This highlights the need for "humane interaction" that considers user autonomy, control, and meaningful participation. Recent research has identified several key requirements for human-centered AI interaction. These requirements include enabling users to question and contest AI outputs [26], providing mechanisms for users to adjust systems based on specific contexts, and establishing clear protocols for responsible AI use. However, major challenges remain in designing interactions that balance automation with meaningful human control, making AI systems more understandable and adaptable to human needs. Better implementing humane interaction has also become a key challenge [3](see Fig. 1.3.) To address this challenge, interaction systems must be "designed responsibly to enhance rather than replace human capabilities" and ensure interactions are "accessible, understandable and trustworthy." [3] Harper et.al. [27] further point out that successful human-AI interaction requires not just system explanations but thoughtful interaction design based on Human-computer-interaction(HCI) principles that enable users and AI systems to work collaboratively. Therefore, developing new interaction methods and interfaces that better align with human-centered AI principles has become increasingly urgent. These developments should focus on enhancing rather than replacing human capabilities, ensuring transparency and user agency, and creating more natural and effective human-AI collaboration. Designing effective human-computer interfaces

to support creative workflows has become especially important, particularly in the context of increasingly prevalent generative AI and large language models.

1.2 Research Objectives

This section defines the three main objectives of the study: analyzing factors affecting prompt effectiveness, developing interactive tools for prompt optimization, and bridging the gap between user intent and the generated images in text-to-image models.

Despite the revolutionary advancements in generative AI and the widespread adoption of Large Language Models (LLMs) in creative workflows, users still face significant challenges in designing effective prompts for text-to-image generation. These challenges include difficulties understanding the complex relationship between prompts and generated outputs, articulating creative intentions clearly, and lacking efficient feedback mechanisms during the iterative process. Such issues make it harder for users to achieve their desired outcomes and significantly increase the complexity of the creative process. While existing methods [20] [21] [28] show some potential in optimizing prompts, they primarily rely on trial-and-error approaches, lack intuitive and user-friendly interactions, and are particularly challenging for non-expert users.

To address these issues, this study proposes a Human-centered framework to improve the efficiency and effectiveness of prompt design. Specifically, the objectives of this research are:

- **Investigate key factors influencing prompt effectiveness:** Systematically analyze how different components of prompts (e.g., structure, keywords, style descriptors) affect the quality and relevance of generated outputs and uncover principles for designing effective prompts.
- **Develop interactive tools for prompt optimization:** Design and evaluate systems that simplify the prompt design process by providing real-time feedback, alternative suggestions, and structured templates, enabling users to optimize prompts efficiently.
- **Bridge the gap between user intentions and model outputs:** Leverage the combination of Large Language Models (LLMs) and Contrastive Language-Image Pre-training(CLIP) [29] to enhance the understanding of user intentions through interpolation and prompt reinforcement techniques. Furthermore, intuitive interaction mechanisms are developed to help users effectively align prompts with their desired outputs while enabling new avenues for creative exploration.

Ultimately, this thesis seeks to help users gain a clearer understanding of the relationship between prompts and generated images, transforming trial-and-error exploration into meaningful and creative adjustments. Through the proposed system, users can generate images that better match their expectations while learning the principles of prompt design, thereby improving the efficiency and quality of creative expression.

1.3 Proposed System

This section introduces PromptNavi, a system designed to simplify prompt optimization through a visual interface and attribute adjustment, supporting users with varying levels of expertise.

Building on the challenges identified in the previous section, this thesis presented PromptNavi, an interactive system designed to assist users in visually exploring and iteratively refining text-to-image generative models. PromptNavi leverages large language models to analyze the user's initial text prompt and refine its attributes through interpolation, generating suggestions that better align with the desired visual outcome. Meanwhile, the system adopts a node-and-connection-based interface, where each node represents an image or a prompt component, and connections indicate relationships between attributes. This structured approach allows users to visually explore and refine prompts more intuitively. And, the system transforms the previously tedious, repetitive, and opaque process of crafting prompts into a dynamic, interactive creative experience. Attribute weights refer to the numerical values that determine the influence of each attribute, such as color, style, or composition, on the generated image. Users can flexibly adjust these weights on the interface and even transfer parts of one prompt to other images for experimentation. They can continually refine input descriptions through real-time interactive output feedback. The "strength" of connections between nodes intuitively indicates how different prompt elements interact, helping users better understand and control the mapping between text descriptions and generated images.

The system employs large language models (such as GPT) for text processing to conduct fine-grained semantic parsing of the user's natural-language input, extracting key visual elements and conceptual tags. At the same time, moderate prompt polishing and simplification can be applied to enhance the stability and quality of generated images. At the generation stage, users can invoke various methods as needed—ranging from a locally deployed Stable Diffusion model (based on diffusion model techniques) to online services like DALL·E—and they may also incorporate any pre-trained models of their choice to meet a variety of resolutions, styles, or application scenarios.

PromptNavi uses the CLIP module for multimodal attribute alignment to embed images and texts in a shared space, enabling bidirectional similarity analysis. On the one hand, this approach evaluates how well the generated image aligns with the target description; on the other hand, it can analyze existing images by comparing them against textual attributes, thereby identifying or correcting any potential mismatches. Through this bidirectional "text-image-attribute" mapping, users can continuously fine-tune prompts or attribute weights during the creative process, achieving precise control over the final visual details and overall effect.

By employing this interactive workflow, PromptNavi guides users in freely discovering and integrating the desired attributes during iterative generation while also effectively excluding or downplaying unwanted elements. The system significantly reduces the cognitive load and time cost of prompt tuning through intuitive attribute manipulation and instant visual feedback, making it readily accessible to beginners. At the same time, PromptNavi provides granular customization options for advanced users, allowing them to tap into the potential of generative models fully. Lowering the barrier to generative AI technology further can attract a broader audience, enabling them to fully realize their creative visions and produce diverse images that match their imaginations. In conclusion, this thesis makes the following contributions:

- PromptNavi is proposed as a node-based visual interface that clarifies the relationships among various prompt elements. It also supports interactive prompt exploration and refinement, making it easier for novice users to generate images from text.
- By leveraging LLM-based attribute interpolation, the system efficiently refines users' prompts while helping them better understand the connection between images and prompts.
- Through text-image similarity, the system provides bidirectional alignment to ensure consistency and offers users a richer selection of attributes.
- By supporting multiple backends (e.g., Stable Diffusion, DALL-E, pre-trained models), addressing the diverse creative needs of different users.
- A user study involving 16 participants demonstrates the effectiveness of PromptNavi and its advantages over commonly used baseline tools.

1.4 Structure

This thesis includes 10 chapters, each complementing the others, collectively focusing on the interactive optimization of generative artificial intelligence (AI) in

Figure 1.4: *PromptNavi* is an AI-assisted image generation and Attribute interpolation system. Users enter a simple prompt (A) that generates an initial image (B). The system then automatically analyzes and extracts visual attributes from the image, categorized into three main groups: Appearance (C), Composition (D), and Atmosphere (E). These attributes can be connected to an empty frame node (F). Once the connections are established, the system can generate a new image (G) that inherits and combines the selected visual characteristics from the original image.

text-to-image generation. It provides a detailed explanation spanning from the research background to system design and evaluation.

Firstly, the Introduction outlines the rapid development of generative AI technology, particularly its extensive applications in the text-to-image generation domain, such as creative design and cultural heritage preservation. However, in practical applications, users often face significant challenges when designing and optimizing prompts, such as discrepancies between generated results and expectations, as well as the complexity and inefficiency of the refinement process. To address these issues, this study proposes a novel interactive system, PromptNavi, and elaborates on its design goals and research significance.

Next, the Related Works chapter reviews current advancements in generative models, user interface designs that support creativity, and prompt engineering. This systematic review analyzes the strengths and limitations of existing methods, further clarifying PromptNavi's research positioning and innovative contributions.

The Preliminary Study chapter focuses on user needs and the pain points of existing tools. Through a combination of interviews and experiments, it identifies significant differences in how various user groups (e.g., novices and advanced users) approach prompt design and optimization. These insights provide clear directions for system design and lay the theoretical groundwork for the design objectives.

The Design and Implementation chapter describes PromptNavi's technical framework and interactive design in detail. By introducing a node-based visual interface, attribute interpolation, and dynamic connection mechanisms, the system transforms the traditional text-based prompt design process into a more intuitive and actionable interaction, enabling users to easily adjust and optimize prompts.

The User Study chapter comprehensively evaluates PromptNavi's practical effectiveness through scientifically designed experiments. Comparative tests with a baseline system demonstrate that PromptNavi significantly reduces cognitive load, improves generation efficiency, and enhances creative exploration. Experimental data and user feedback substantiate the system's effectiveness.

The Results chapter further analyzes experimental data, detailing PromptNavi's performance across multiple dimensions, such as user experience, task completion efficiency, and the quality of generated results. It also highlights specific behaviors and user feedback during their interaction with the system.

The Discussion chapter analyzes the findings deeply, exploring PromptNavi's design principles, application scenarios, and limitations. It suggests future research directions, such as better support for complex scenarios in prompt design and further enhancements to the system's scalability and applicability.

The Limitations chapter focuses on the shortcomings of this research, such as the relatively weak traceability of prompt optimization history and the need to improve the interface's operational efficiency in highly complex scenarios.

Finally, the Conclusion and Future Work summarizes this research's core contributions to interactive optimization for generative AI and provides an outlook on possible future research directions.

Chapter 2

Related Works

This chapter reviews advancements in text-to-image generative models, creative support interfaces, and prompt engineering and identifies issues that PromptNavi aims to address.

2.1 Text-to-Image Generative Models

This section examines the evolution of text-to-image models, from early GAN-based methods to recent diffusion models like Stable Diffusion, highlighting usability challenges for non-experts.

Text-to-image generative models create images based on textual descriptions by combining natural language processing and computer vision techniques. Early methods mainly used Generative Adversarial Networks (GANs) to map text to images. For example, AlignDraw [30] extended the RNN-based Draw model by using a bidirectional attention RNN to guide image generation. Reed et al. [31] introduced a conditional GAN model, which could generate simple images from text but provided limited control over user input.

Recent advances in deep learning, especially Transformer architectures and large-scale pre-trained models [4], have driven significant progress in text-to-image generation. OpenAI's DALL·E series [8] used large-scale datasets, such as LAION-5B and other web-scraped image-text pairs, along with Transformer models to generate high-quality images from complex text descriptions. However, these models often lack direct feedback to reflect user intent, requiring users to refine prompts to achieve desired results iteratively.

Diffusion models have introduced new possibilities in this field [32]. GLIDE [33] and Stable Diffusion can generate high-quality, detailed images through a gradual denoising process. While these models improve image quality, they also increase the complexity of user interactions due to their reliance on carefully crafted

prompts.

Multimodal pre-trained models like CLIP [29] further enhance text-to-image generation by aligning text and image representations. Trained on large-scale datasets, CLIP learns rich semantic relationships and is widely used to guide image generation. However, its usability for non-expert users remains limited.

Despite these advancements, prompt design remains challenging due to its complexity and the trial-and-error nature of refining inputs. PromptNavi combines Large Language Models (LLMs) and CLIP to develop an interactive system focused on prompt creation and user interaction.

2.2 User Interfaces for Creativity Support

This section explores interactive interfaces that aid creativity, noting the limitations of current tools in supporting iterative and user-friendly prompt refinement.

For generative AI, user interfaces (UIs) are a crucial bridge between complex machine-learning models and end-users, especially in creative workflows. Effective UI design can lower technical barriers, enabling users to leverage better generative AI for tasks such as image creation, music composition, and 3D modeling. For instance, the user interface of Stable Diffusion is often implemented through third-party tools, allowing users to generate high-quality images simply by entering prompts. However, these interfaces primarily support essential input and output operations, lack advanced features for fine-grained editing of the generated results, and often need to engage in brute-force trial and error with the text prompt when the resulting quality is poor [34]. Similarly, ChatGPT [35], as a conversational generative tool, enables users to produce content via natural language inputs. While convenient, these tools lack iterative mechanisms, limiting their ability to support complex creative tasks effectively.

In recent years, user interface design has evolved from traditional static controls to more complex interactive paradigms. For example, LayoutLLM-T2I [36]uses large language models to convert prompts into layout vectors through a feedback-driven induction mechanism, enabling users to manipulate image generation by adjusting layout frames. Similarly, PromptPaint [37] adopts paint medium-like interactions, allowing the users to transform text prompts into flexible vector forms through prompt mixing and directional prompts, allowing them to explore image generation, like blending colors. These interactive interfaces significantly enhance users' understanding of the relationship between inputs and outputs, particularly in tasks with well-defined parameters. However, similar research still relies on detailed text prompts in more open-ended or complex tasks, requiring users to have convenient input methods or assistive tools to support higher-quality generation.

Existing generative AI interfaces exhibit significant limitations in such tasks.

Most interfaces lack fine-grained control features, forcing users to adjust input prompts through trial-and-error methods. The trial-and-error method of manually refining prompts is inefficient, particularly for non-expert users, making it challenging to achieve desired results quickly. Many researchers have proposed solutions to address this issue. For instance, PromptCharm [38] proposed prompt optimization and feedback mechanisms to help users automatically refine prompts and visualize the model's attention to specific keywords, enabling more precise adjustments. Similarly, Promptify [28] adopts an interactive prompt exploration approach, using GPT-generated suggestions and image clustering attributes to guide users in iteratively improving prompts and managing extensive collections of generated images. However, the complex interaction designs chosen by such studies may increase the learning curve, especially for users with no prior experience, who may require more time to adapt to multi-step workflows.

To address these issues, research in prompt engineering and human-AI collaboration can inform the development of more human-centered interfaces. For example, integrating LLMs (e.g., GPT series) with multimodal models (e.g., CLIP) could enhance the connection between prompt design and content generation, allowing users to grasp how their inputs influence the results intuitively. Such a design approach can provide a more seamless and efficient experience for complex creative tasks.

2.3 Prompt Engineering for Image Generation

This section discusses the evolution of prompt engineering from manual input to systematic optimization and underscores the need for interactive and intuitive tools like PromptNavi.

Prompt Engineering has evolved from experience-driven text tuning into a systematic process integrating linguistics, semantic modeling, and tool-assisted optimization. Its goal is to maximize generation quality, controllability, and interpretability without altering model parameters [39].

For instance, Brown et al. [40] demonstrate that large-scale language models achieve few-shot learning using task descriptions and minimal examples, excelling in translation and Q&A tasks. Wu et al. [41] proposed chaining operations, breaking complex tasks into manageable subtasks to improve transparency and user interaction. Similarly, Wen et al. [42] introduced a gradient-based method to optimize hard and soft prompts, achieving efficient and interpretable results while bypassing content filters.

For image generation AI, prompts have become the primary user interaction tool, enabling users to generate highly customized visual content through precise descriptions, style selection, and multi-modal integration.

For example, Hertz et al. [43] proposed a Prompt-to-Prompt image editing method that uses cross-attention mechanisms to enable text-based image editing. This approach allows localized and global modifications by adjusting the prompt while preserving the original image structure. Cao et al. [44] introduced the BeautifulPrompt, which uses reinforcement learning with visual feedback to optimize low-quality user inputs. This helps generate prompts that better guide diffusion models to produce high-quality images. These methods still rely on traditional manual input from users, which requires a certain level of prompt-generation knowledge. In addition to the methods [20] mentioned in Chapter 1, researchers have turned to interactive approaches that simplify user prompt generation. This also avoids the constraints of traditional end-to-end methods [13], making the relationship between prompt generation and image output more controllable.

In Prompirit, Kim et al. [45]proposed an automatic prompt engineering system that enhances AI-generated artwork by incorporating emotion labels and style modifiers to improve emotional expressiveness and aesthetic quality. While the system achieves fine-grained control over semantic and visual styles, it relies on a preset labeling framework, limiting dynamic adjustments and personalized creative expression.

Similarly, Feng et al. [46]introduce an interactive system that recommends keywords and visualizes their impact to help users refine prompts for text-to-image generation. The system enables intuitive exploration of keyword-image relationships by leveraging a large-scale prompt-image dataset. However, it lacks real-time feedback and dynamic attribute controls, relying heavily on user judgment and repetitive trial-and-error processes.

Through these studies, it is observed that users' needs in image generation require a balance between creative expression and the exploration process. Most existing research focuses on improving the final output, often overlooking the creative insights and more profound understanding users develop through iterative exploration. Furthermore, users may have higher demands regarding image composition, visual depth, and stylistic presentation, with each exploration prompting them to refine or expand their expectations and creative goals.

Therefore, the goal is to identify a balance point that ensures each user exploration delivers new value, deepens their understanding of the relationship between prompts and generated images, and creates a more controllable and intuitive exploration experience. In this work, these core user needs are validated through preliminary research, and PromptNavi is proposed as an interactive prompt optimization platform that combines real-time visual feedback with attribute interpolation mechanisms. PromptNavi empowers users to generate images efficiently, explore creative ideas, and achieve personalized expression throughout the iterative process.

Chapter 3

Preliminary Knowledge

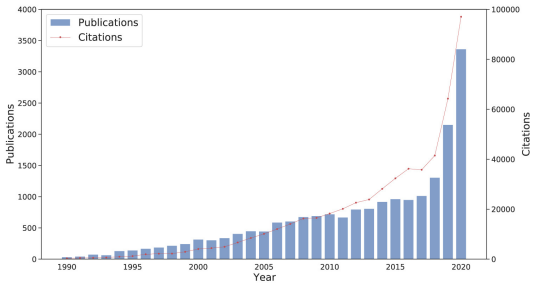
This section provides an overview of the fundamental concepts of the Pre-trained Model, CLIP Models, and the research questionnaires used in this study, including the NASA-TLX and SUS scale.

3.1 Per-trained Model

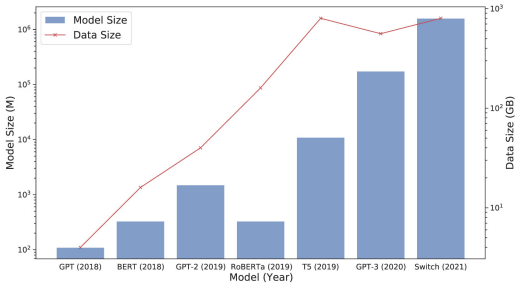
Pre-trained Models (PTMs) are a type of deep learning model that is pre-trained on large-scale datasets and then fine-tuned on specific downstream tasks. During the pre-training phase, they learn general knowledge representations, which are transferred to specific tasks through fine-tuning [47] [4]. From a developmental perspective, the evolution of PTMs has gone through several significant stages. In the earliest phase, pre-training was mainly based on transfer learning theory, using attribute and parameter transfer to achieve cross-task knowledge transfer. This stage achieved remarkable success in computer vision, exemplified by CNN models such as AlexNet [48] and VGG [49], which were pre-trained on ImageNet [50].

The real breakthrough occurred in 2018 when the emergence of BERT [51] and GPT [35] based on the Transformer architecture ushered PTMs into a new era and introduced various AI products to the public. These models can learn context-dependent representations and thus better understand language's semantic and syntactic structure. Their size expanded from the early millions of parameters to billions(see Fig. 3.1). By 2020, large-scale language models represented by GPT-3 showcased astonishing capabilities. Trained on 560 GB of data and 10,000 GPUs, GPT-3 demonstrated strong language understanding and generation abilities and exhibited human-like few-shot learning capabilities—able to produce user-satisfying outputs with just a few simple prompt examples [40](see Fig. 3.2).

PTMs have achieved such remarkable success primarily due to their unique advantages. First, PTMs can learn a wealth of knowledge—including linguistic



(a) The number of publications on "language models" and their citations in recent years.



(b) The model size and data size applied by recent NLP PTMs. A base-10 log scale is used for the figure.

Figure 3.1: (a) shows the number of publications with the keyword "language model" and their citations in different years. (b) shows the parameter size of large-scale PTMs for Natural language processing(NLP) tasks, and the pre-training data size increases by 10 times per year. From these figures, it can be observed that, after 2018, when large-scale NLP PTMs began to be explored, increasing efforts have been devoted to this field, and the model size and the data size used by PTMs have also grown.(adapted from [4])

knowledge, world knowledge, and common sense [52] from large-scale unlabeled data. This knowledge can be effectively transferred to downstream tasks, significantly reducing the need for labeled data. Second, a single pre-trained model can be adapted to various downstream tasks through simple fine-tuning, significantly lowering the cost of training a new model from scratch. Moreover, in many tasks, PTM-based methods have surpassed traditional approaches, with some reaching or exceeding human-level performance.

In this thesis, the proposed system did not train or fine-tune any models for image generation. Instead, pre-trained models were leveraged due to their broad applicability and convenience. Adopting PTMs as the backbone for downstream tasks instead of training models from scratch has become common in the AI community. This consensus is grounded in PTMs' significant performance advantages and paradigm-shifting impact on AI system development. By leveraging a pre-trained–fine-tuning approach, AI systems can be developed more efficiently, enabling more applications to benefit from large-scale pre-trained models.

3.2 CLIP Models

CLIP (Contrastive Language-Image Pre-training) is a visual language model proposed by OpenAI [35]. It employs a pre-trained method based on contrastive learning that combines a visual coder and a text encoder to learn multimodal representations by aligning semantic relations between image and text descriptions.

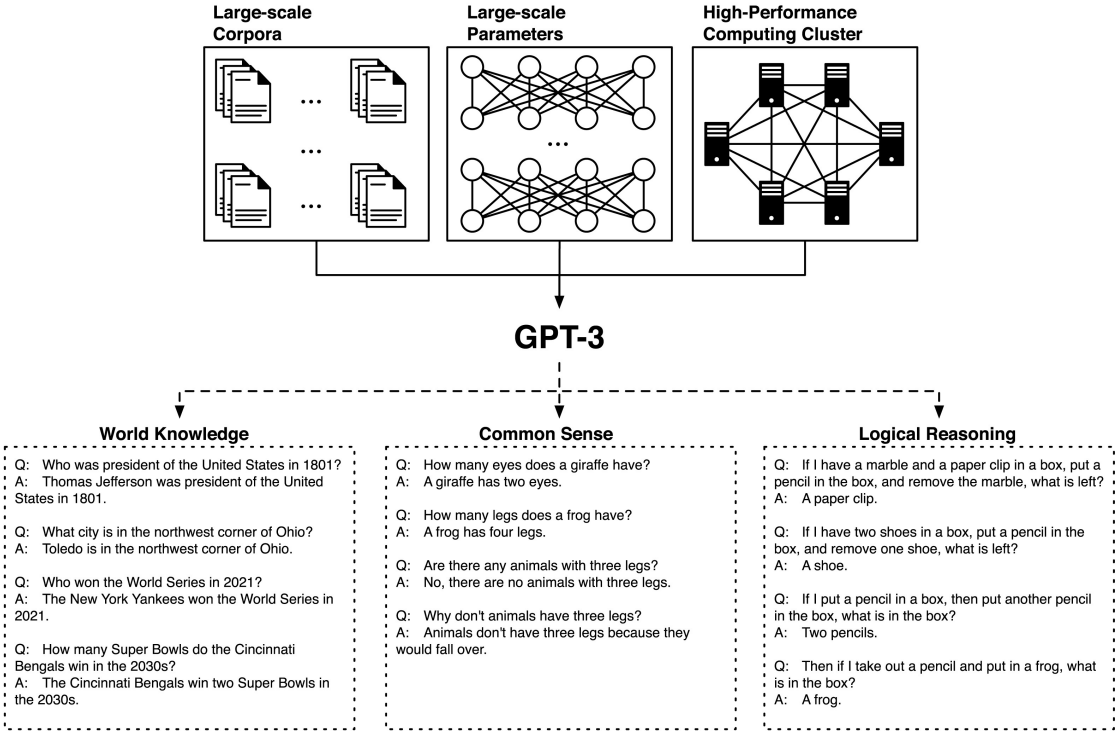


Figure 3.2: GPT-3, with 175 billion parameters, uses 560 GB data and 10,000 GPUs for its training. It has shown the ability to learn world knowledge, common sense, and logical reasoning. (adapted from [4])

CLIP is trained on 400 million pairs of noisy image-text data collected from the Internet, and does not rely on manually labeled large-scale categorization or target detection datasets. The core idea is to link vision and language through a contrast learning approach. The model encodes images and text by a visual encoder and a text encoder respectively, and uses a contrast loss function to maximize the similarity of paired images and text and minimize the similarity of mismatched pairs. As shown in Fig. 3.3, unlike traditional models, the training data of CLIP is not derived from expensive and restricted manual annotation, but is based on 400 million pairs of image-text descriptions crawled from the Internet. This design breaks through the limitations of manual annotation and allows the model to utilize rich and diverse natural language supervision, thus learning a wide range of visual concepts and linguistic expressions. CLIP is distinguished by its powerful zero-shot learning [53] capability. While traditional models often require additional fine-tuning for each task, CLIP can rely only on generalized knowledge learned from pre-training to make direct inferences in new tasks. For example, in the ImageNet classification task, CLIP has outperformed many fine-tuned models

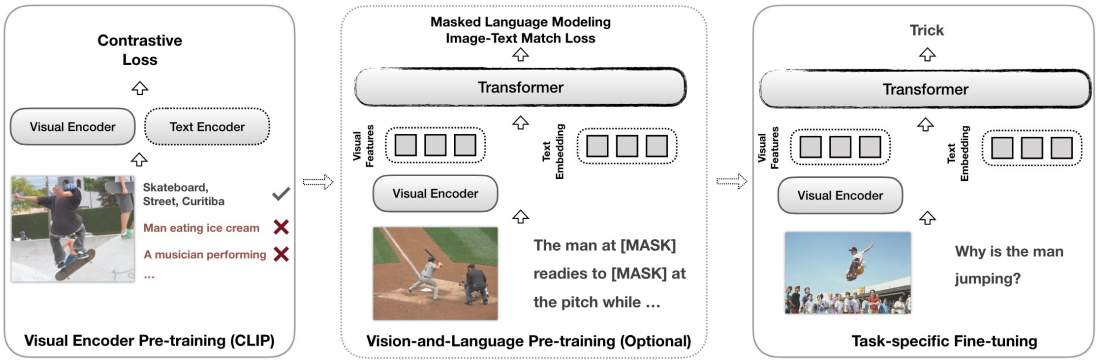


Figure 3.3: The training process of a Visual and language(V&L) model typically consists of three steps: 1) visual encoder pre-training, 2) vision-and-language pre-training (optional), and 3) task-specific fine-tuning. In previous V&L models, visual encoder pre-training requires human-annotated vision datasets, which are hard to scale.(adapted from [5])

in a zero-sample setting. This capability stems from CLIP's ability to unify vision and language into a shared semantic space, allowing it to process unseen tasks and data naturally. As shown in Fig. 3.4, CLIP's visual coder demonstrates significant performance advantages over traditional region-feature [54] or grid-feature [6] [55] coders, especially in tasks that require complex cross-modal reasoning.

CLIP's strengths are its training efficiency and flexibility in data scaling. Its shallow-interaction design allows the visual and text encoders to process input data independently, significantly reducing computational cost [29]. In addition, CLIP does not rely on predefined category labels. It can learn generic representations beyond specific classification tasks by aligning the semantic relationships between image and text descriptions. This unconstrained data utilization allows CLIP to show greater adaptability when facing complex and dynamic real-world tasks. Nevertheless, CLIP is still limited to specific tasks requiring deep cross-modal reasoning. For example, complex reasoning problems in visual quizzing and understanding fine-grained semantics must be combined with deeper multimodal interaction design [5]. In addition, CLIP's lack of localization capability limits its effectiveness in tasks such as target detection. Future research directions could include exploring the combination with deep interaction models, optimizing the visual feature representation of CLIP, and extending the diversity of training data to improve its performance and adaptability further.

Overall, CLIP's proposal brings a breakthrough in multimodal learning. By integrating semantic representations of vision and language, CLIP extends the boundaries of AI in the multimodal domain and provides strong support for zero-

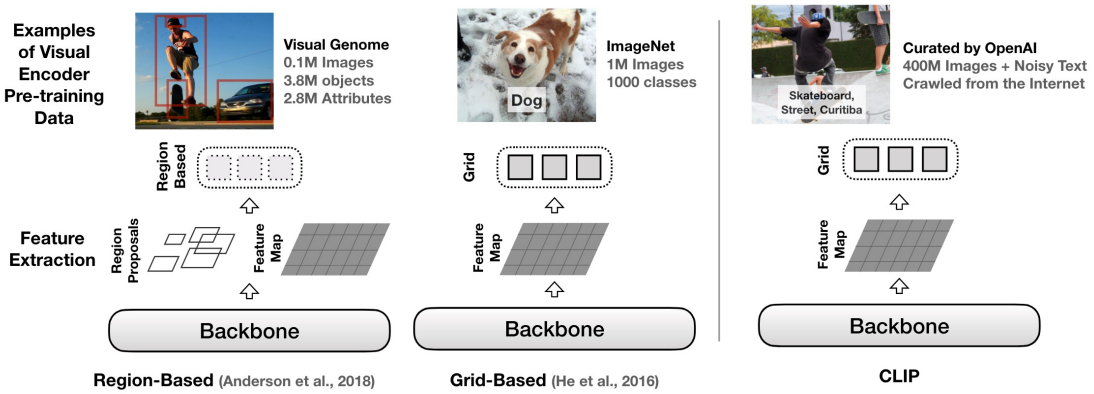


Figure 3.4: CLIP versus other visual encoders. Region-based methods are trained on object detection data. Previous work uses either image classification [6] or detection data for grid-based methods. However, CLIP requires only aligned text.(adapted from [5])

sample learning and the realization of generalized intelligence. Its success enlightens future research possibilities on building more generalized and intelligent multimodal models with more extensive and diverse data and more efficient training methods.

Chapter 4

Preliminary Study

This chapter details the initial user interviews and experiments conducted to understand user needs and challenges in prompt design, providing the foundation for system development.

A preliminary study was conducted through formative interviews to better understand users' needs and limitations when generating desired images using prompt-based image generation AI. The focus was on users' interaction and exploration processes with the AI system, particularly how they compose and iterate text prompts to conceptualize visuals and respond to the generated outputs.

4.1 Participants

This section describes the demographics and experience levels of the study participants, ranging from novice to advanced users, to capture diverse perspectives.

Eight graduate students (1 female, 7 male; ages 23–28) were interviewed. Based on their experience and frequency of using image generation tools, the participants were categorized into novice, intermediate, and advanced users. The participants self-reported their familiarity with image generation as follows:

- **Novice users** (P1, P2): These participants have used image generation tools for less than 6 months, with an average usage frequency of no more than 5 times per week, and possess limited understanding of the related technologies.
- **Intermediate users** (P6, P7): They have used the tools for 6 months to 1 year, with an average usage frequency of 5–10 times per week, and have some practical experience.

- **Advanced users** (P3, P4, P5, P8): These participants have used the tools for over 1 year, with an average usage frequency exceeding 10 times per week, and have related research experience.

Informed consent was obtained from all participants. All participants agreed to have their interview content recorded and their results published anonymously.

4.2 Formative Interviews

This section summarizes insights from user interviews, highlighting differences in interaction styles, iteration processes, and challenges faced by users of varying expertise levels.

4.2.1 Differences in Interaction Styles Across User Skill Levels

Users of different skill levels exhibit notably different interaction styles when interacting with text-to-image generation models. Novice users often prefer to provide minimal descriptions and quickly obtain a roughly acceptable outcome, investing little time and effort in prompt design. For example, P6 (intermediate user) states: *"I adhere to a 'keep it simple' principle—usually just one concise sentence, without piling on too many attributes or details."* Such users treat the generation model as a "black box," hoping to achieve workable outputs with minimal linguistic complexity. In contrast, advanced users embrace more complex prompts and multimodal inputs to achieve richer control. For instance, P8 (advanced user) emphasizes: *"I sometime provide a simple sketch alongside prompts... This helps ensure the generated image aligns more closely with my initial vision."* This approach shows that advanced users understand the importance of prompt design and proactively incorporate sketches and other visual aids to enhance control over the resulting image. P5 similarly mention using various auxiliary inputs—such as depth maps and sketches—to enrich prompts, indicating that more experienced users are willing to explore multiple information channels.

This comparison reflects differences in both user experience and motivation: novice users favor a "quick and rough" approach and lack the will or capacity to explore the full diversity of prompt design, whereas advanced users engage in "fine and detailed" iteration, continuously refining prompts and inputs to achieve higher-quality outputs. As a result, novice users struggle to produce complex scene descriptions or fine-grained control. In contrast, intermediate and advanced users must invest substantial time in crafting prompts that lead to ideal outcomes.

4.2.2 From One-Off Attempts by Novices to Iterative Refinement by Advanced Users

Regarding interaction processes, novices usually try a prompt once and make only minor tweaks, lacking the drive or methodology for deeper iterative refinement. P6, for example, after trying a straightforward sentence, only makes slight keyword adjustments before stopping. By contrast, intermediate and advanced users favor multi-round iteration and gradual optimization. P1 explains: *"I typically begin with an elementary and concise description... then add more specific elements to refine the output,"* illustrating a step-by-step approach. P3 further shares the importance of fine-grained iteration: *"Then, I gradually add details about the background environment, the character's posture, and the overall artistic style."* showing that experienced users regard prompt crafting as a continuous trial-and-error process, steadily approaching their ideal image. P7 similarly notes: *"I begin with core keywords or phrases and then gradually add more detailed, specific terms."*

This feedback indicates that for experienced users, each generation provides feedback, enabling them to add elements and fine-tune descriptions systematically. Although there is no direct mention of truly infinite loops or entirely diverging outcomes, P1, P3, and P8's statements imply that users often must repeatedly experiment and adapt to model outputs, approaching the desired result incrementally. This hints at the current tools' lack of practical guidance or stopping conditions, as users might get caught in repeated trial and error without a direct path to quickly achieving a satisfactory output.

4.2.3 From "Lacking Expression Strategies" to "Tool Limitations in Complex Multi-Element Control"

The difficulties users face vary across different skill levels. Novice users often struggle to articulate their needs, not knowing how to refine prompts when initial descriptions fail to produce the desired effect. When the basic prompt doesn't meet their expectations, they might feel stuck, unsure which keywords can more precisely guide the AI. Even users with some experience, as P1 notes, must "try-observe-try again" by examining initial results before deciding which elements need reinforcement.

Intermediate users, though willing to refine prompts, still find it challenging to achieve ideal outcomes in complex scenes. P7 describes using a basic English description ("Old man, long hair, with glasses.") and resorting to translations and ChatGPT refinements when unsatisfied, but improvements remain limited. Advanced users (such as P5) point out: *"Even with precise text prompts and sketch-based assistance, inconsistencies... can still occur for multi-object generation."* Fine-grained control in multi-object or complex scenarios remains challenging even

with accurate descriptions and auxiliary cues. This reveals a progression of difficulties: novices struggle to "get started," while advanced users find "fine-tuning" complex scenes difficult. Thus, Supporting all these users becomes a central issue: providing more precise controls and guidance that reduce these hurdles.

4.2.4 Refining Unsatisfactory Results Through Repeated Iteration

When trying to improve unsatisfactory outcomes, users across all skill levels inevitably engage in multiple rounds of iteration and trial-and-error. Novices typically only make slight keyword changes or settle for suboptimal results. As P6 says: *"If the initial output doesn't quite match what I had in mind, I tweak that same prompt slightly,"* indicating a lack of effective strategies to handle discrepancies. Intermediate and advanced users have more nuanced responses. P1 makes targeted keyword adjustments to approach the ideal image, and P3 introduces negative prompts to reinforce generation direction: *"...while introducing some negative prompt words to reinforce the direction of generation."* While these advanced techniques can improve iteration efficiency, they do not entirely circumvent repeated trial and error. Users lack more direct and intuitive tools to quickly pinpoint and correct issues without continuously relying on incremental textual refinements.

4.2.5 From Intuitive Attribute Control to Deep and Multimodal Assistance to Meet Various Complexity Needs

User expectations for tool improvements range from simple, intuitive features to deep, fine-grained, and multimodal controls. Novice users want more straightforward ways to reduce trial and error. For example, P6 hopes: *"If the image could be updated in real-time as I manipulate those sliders."* thus seeking visual controls that enable quick comprehension and adjustment without extensive linguistic descriptions. Intermediate users, such as P7, leverage ControlNet [56] and sketches to complement text-based prompts, making it easier to fine-tune style and content. Advanced users demand even greater control and refinement options. P8 suggests using percentage sliders for fine-tuning parameters: *"A more intuitive method—like using percentage sliders—would make fine-tuning more transparent."* and wants keyword classification, examples, and structured management to simplify creative exploration in complex scenarios. P5 emphasizes the value of high-quality preset prompts and intelligent suggestions. These expectations indicate that existing tools still fall short in areas like localized editing, granular control, multimodal interactions, and intelligent support, failing to fully meet users' needs for rapid and accurate image mastery.

4.2.6 Findings

Overall, users at different skill levels show a clear stratification in interaction styles, iterative processes, challenges, coping strategies, and desired functionalities. Novices favored simple input and quick initial results but faced barriers in deeper refinement and complex control; intermediate users attempted to enhance results by adding descriptions and using external aids but remained constrained in complex scenarios; advanced users possessed more mature optimization strategies and multimodal input capabilities yet are limited by current tools' shortcomings in fine-grained control, diverse input, and intuitive adjustments. P1 and P3's emphasis on stepwise iteration, P5's identification of multi-object scenario difficulties, and P8's expectation for fine-grained adjustments and structured prompt management all offer valuable insights for tool design.

Building on these findings, PromptNavi can address the needs of different user levels through a range of assistance features—such as automatic keyword suggestions, attribute interpolation visualization, negative prompt management, and multimodal support. Such a layered approach lowers the entry barrier for novices, enriches controllability for intermediate users, and provides advanced users with efficient, flexible, and deeply customizable creative tools. While current systems still have limitations and have not achieved fully controllable and efficient image generation in all complex scenarios, these limitations highlight areas for future improvement and innovation. By evolving to meet these diverse and progressively complex needs, the creative potential of text-to-image generation can be better unlocked for users at every skill level.

4.3 Preliminary Experiment

This section discusses a preliminary experiment in which participants used existing text-to-image generative AI tools to identify common issues and pain points in prompt design and iteration.

A preliminary experiment was conducted with 8 participants from different user groups, all with a background in HCI research. The purpose was to gain insights into how users explore and utilize generative AI and better understand its usage scenarios. In this experiment, participants used a traditional generative AI, the commonly used Stable Diffusion, and an open-source UI, Automatic1111 [57]. This UI provides basic visual functionality for image generation but does not offer writing assistance, encouraging users to focus more on crafting and iterating their prompts. And also provided a clear perspective for us to analyze their exploration patterns.

During the experiment, each participant started with a simple prompt, ob-

served the generated image, and then manually modified and iterated their prompts to create subsequent images. After four iterations, the prompts and images generated by the participants were analyzed. Brief interviews were also conducted to better understand the typical scenarios and behavioral patterns exhibited by users from different groups when interacting with generative AI.

The result was taken as the scenario sample, as shown in Fig. 4.1; in scenario 1, The user has a clear goal. Their iteration strategy involves examining the image and making a planned rewrite. Consequently, the initial and iterative prompts are rarely retained in the final version. In scenario 2, The user starts with a vague concept and is unclear about the details. They prefer to establish an overall direction first and then iteratively add details. Hence, the initial and iterative prompts are primarily preserved in the final version. In scenario 3, The user is exploring without a specific purpose, lacking clarity in direction and details. They refine their needs through observation; therefore, the initial and iterative prompts are mainly kept in the final version.

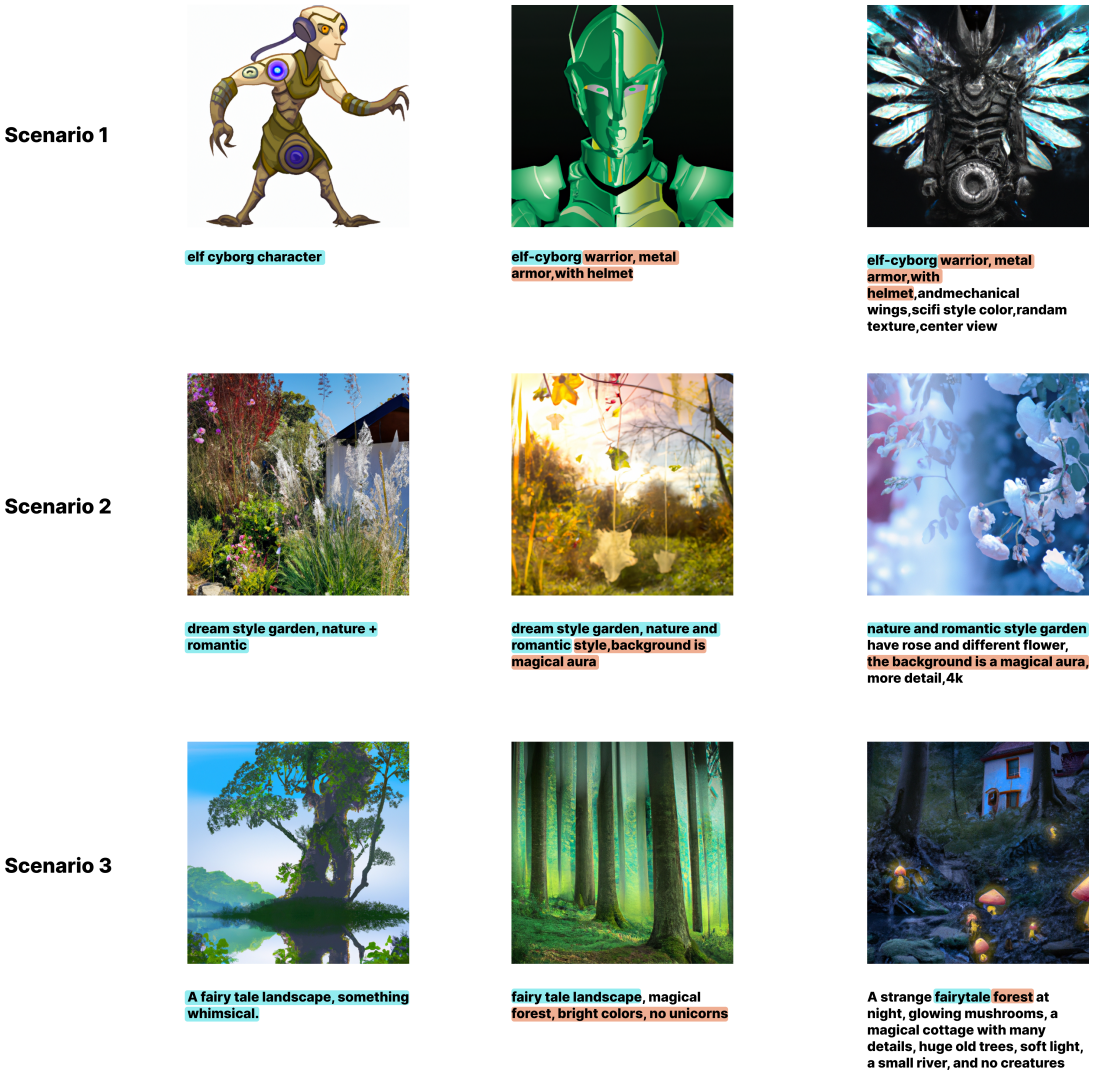


Figure 4.1: Examples of the generated images from the preliminary experiment.

Chapter 5

Design Goals

This chapter outlines the key scenarios and research questions that shaped PromptNavi's design, with specific hypotheses guiding its development.

5.1 Scenario

This section identifies three typical user scenarios: repetitive iteration, gradual refinement, and open-ended exploration, which inform the system's feature design. Based on the preliminary study, the following three typical user scenarios were identified:

- **Scenario A: Repetitive Effort in Iterative Optimization**

  When users need to generate multiple images with consistent styles but different content (e.g., a series of advertising images or design sketches), they typically aim to start from an initial template and make specific adjustments for each image. However, traditional generative AI user interfaces, such as Stable Diffusion WebUI and DALL·E's standard interface, present significant challenges in this process. Each iteration requires users to rewrite prompts from scratch, making it difficult to reuse prior knowledge and style settings. This repetitive effort wastes time and energy and risks compromising the overall consistency of the image set.

- **Scenario B: Gradual Refinement of Complex Concept**

  When users have a vague conceptual idea in mind, with details still undefined and difficult to articulate, they often begin by inputting a general prompt and then gradually add details to refine the concept. However, traditional generative AI user interfaces struggle to support such tasks effectively. They heavily rely on users to express their needs precisely through text, which

makes it challenging to translate abstract or ambiguous ideas into suitable prompts. Furthermore, the lack of layered control over key elements of the generated image prevents users from independently optimizing specific parts, often leading to results that deviate from their initial vision.

- **Scenario C: Open-Ended Exploratory Generation**

  When users lack a specific creative goal and instead aim to explore various possibilities and find inspiration through generative AI, they typically input broad or abstract prompts (e.g., "futuristic city" or "dreamlike landscape") and review the resulting images. Such open-ended exploration relies heavily on users' ability to adjust and respond to the generated content. However, traditional generative AI user interfaces face notable limitations in this context: they often produce highly random outputs in response to broad prompts, which may stray far from the user's area of interest. Additionally, users lack practical tools to control or steer the generation process, making it difficult to discover relevant and inspiring creative directions efficiently.

5.2 Research Questions and Hypothesis

Preliminary interviews revealed that novice, intermediate, or advanced users engage in iterative trial-and-error processes and prompt adjustments when using generative AI to create images, aiming to achieve their desired results. However, preliminary research also indicates that, in most cases, these attempts do not translate into actionable insights that support future explorations. This often leads users into blind and ineffective trial-and-error loops when seeking to generate their ideal images. Therefore, the research question is: **If users' trial-and-error cycles can be effectively utilized, could this better assist them in completing their tasks?**

To address the research question, the following hypotheses are proposed:

- **H1:** The proposed system records and visualizes users' trial-and-error processes, enabling users to understand better the relationship between generated outputs and prompts, thereby reducing ineffective trials and improving generation efficiency.

- **H2:** The proposed system's prompt interpolation functionality helps users reduce the burden of repeatedly drafting prompts.

- **H3:** The proposed system's prompt interpolation functionality allows users to complete their tasks while generating prompts they are satisfied with.

- **H4:** The proposed system's design, based on connections and node visualization, is more practical and usable than traditional generative AI user interfaces.
- **H5:** The proposed system is more effective than traditional generative AI interfaces in helping users achieve their artistic visions and goals.

Chapter 6

System Design

This chapter describes the technical framework and interactive design of PromptNavi, detailing its key components, such as the control panel, image nodes, and attribute interpolation. PromptNavi is an interactive AI-assisted image generation

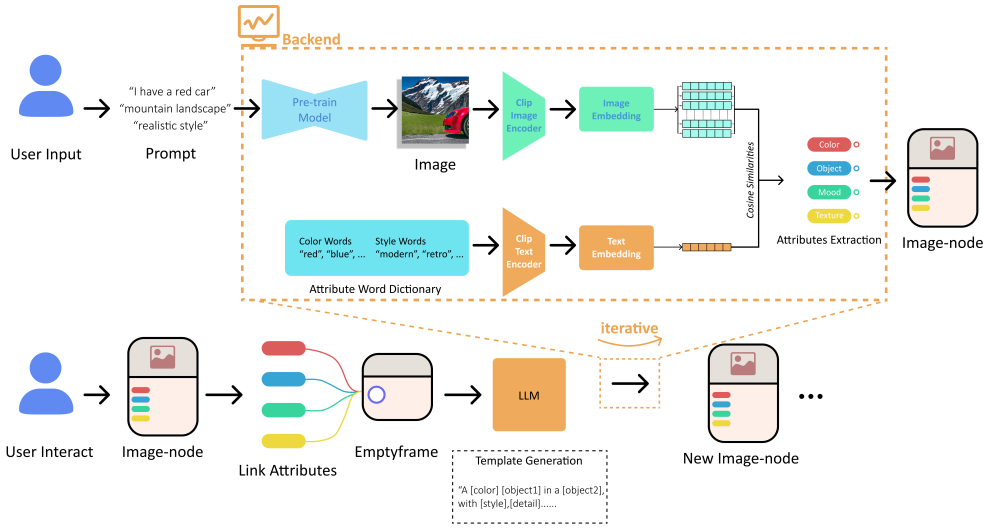


Figure 6.1: Framework of promptNavi. The PromptNavi framework processes user prompts through a pre-trained model and CLIP to extract visual attributes organized in image nodes. Attributes are linked to an empty frame, and the system uses LLMs to generate refined prompts for iterative image generation.

and attribute interpolation system(see Fig. 6.1). Built with a Flask Python backend and vanilla JavaScript frontend architecture, the system provides an intuitive workspace interface. The main interface has two major components: a control panel and a canvas. The control panel integrates prompt input for image generation, model selection (supporting DALL-E 2/3, Stable Diffusion, and pre-trained

custom models), and parameter settings for image size and quality.(see Fig. 1.4A Fig.6.2A)

The system's core interface is a dynamic canvas where generated images are presented as nodes(see Fig. 1.4). As shown in Fig. 6.4, each node contains collapsible panels for attributes and prompts, with the system automatically analyzing visual attributes, including color, style, composition, lighting, mood, object, perspective, detail, and texture attributes. As shown in Fig. 6.5, these attributes can be interconnected through interactive connection points, allowing users to establish attribute relationships between different nodes(see Fig. 6.6) through drag-and-drop operations and adjust the influence of each attribute through weight controls(see Fig. 6.7). The system employs color-coding based on attribute types to provide visual feedback and supports attribute folding and group management to handle complex connection relationships.

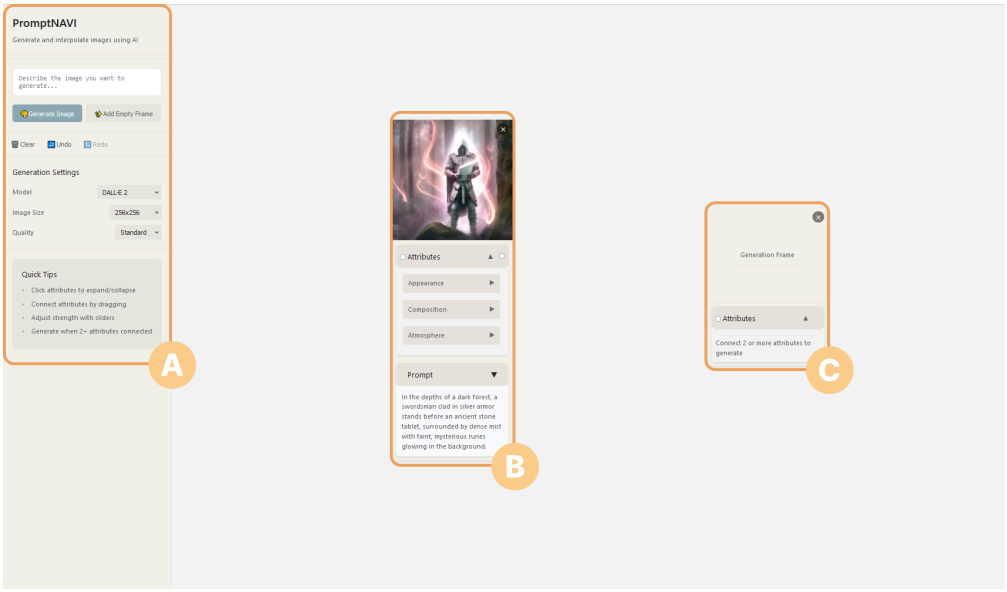


Figure 6.2: Overview of *PromptNavi* structure including a control panel (A), image-node with attributes and prompt (B), and empty frame (C) inside the canvas.

6.1 Control Panel

As shown in Fig. 6.2(A), The main UI interface is illustrated in Fig. 6.2. PromptNavi includes a basic text-to-image control panel within the system, which collects user input and provides straightforward interactive prompts (see all UI components

in Fig. 1.4). This panel can accommodate various models, including online, locally pre-trained, and user-trained. Users can select the desired model and image size on this interface and then input a text description or a base image to generate a resulting image displayed directly on the canvas. This allows users to quickly assess the initial output and decide on their next steps—such as entering new descriptions, switching models, or adding more attributes for further exploration.

6.2 Attributes Extraction

The system's canvas represents each image as an Image node rather than a static image tag. This node-based approach offers greater flexibility for subsequent compositing and iterative processes: the node contains not only the image URL but also stores its associated text prompt and attribute information derived from the backend model (e.g., color, style, or object details). From a user's perspective, this design allows each image node to be freely dragged, expanded, or collapsed on the interface, making it easier to organize and manage multiple images visually.

When processing both the text prompt and the image data, the backend first utilizes a language model (GPT-4o-mini API) or a text-analysis module to parse and extract various keywords and concepts from the prompt (assigning appropriate weights to elements like "blue sky" or "modern style"). Simultaneously, the image is fed into a multimodal model (e.g., CLIP) to generate attribute vectors and detect potential color themes, styles, and object Categories. The system then maps the text and image into a common latent space or applies an attention mechanism to align their concepts: if a particular attribute appears in both the text and the image, their confidences are merged; otherwise, it is tagged as an attribute from either the "text prompt only" or the "image detection only." In the end, these multimodal details are compiled into a structured attributes set—such as

```
{ color: { \blue": 0.8 }, style: { \abstract": 0.6 } }
```

—and returned in JSON format. The front end then visualizes these attributes as labels or "attribute points" in the corresponding Image Node, allowing users to inspect and manipulate each image's attributes with fine granularity.

By extracting and displaying attributes in this manner, the high-dimensional characteristics behind each image are distilled into a handful of human-readable labels, enabling users to flexibly combine or edit nodes according to their creative goals without understanding the underlying multimodal inference. This label-based presentation also helps users perceive how the text prompt correlates with the image's attributes or how modifying specific descriptions might affect the generated content, thereby enhancing the overall controllability and transparency of the creative and iterative workflow.

Figure 6.3: Different values will directly affect the thickness of the line. Linked attributes will appear in the empty frame as "attribute points".

6.3 Attributes Categories

Since users create prompts based on describing images, the attribute categorization design is rooted in practical applications. It draws inspiration from traditional photography principles while addressing the needs of image generation [58]. This system established three main categories: Appearance, Composition, and Atmosphere. This classification provides precise attribute dimensions for generating prompts and allows users to intuitively understand and adjust the relationship between prompts and the resulting images through specific attribute parameters.

The appearance category focuses on an image's visual elements, including color, style, and object. These attributes directly influence the image's overall visual impression and aesthetic appeal. For instance, adjusting color can shape the image's overall tone, optimizing the style can help generate images that meet specific artistic requirements, and emphasizing objects can highlight key content.

The composition category centers on an image's spatial structure and layout, covering composition, perspective, and detail. By modifying these attributes, users can control the arrangement of elements within the image to adhere to visual aesthetic principles. For example, changes in perspective can enhance spatial depth, while refining details can improve the image's richness and layering.

The atmosphere category pertains to an image's overall emotional expression and environmental characteristics, including lighting, mood, and texture. These attributes determine the visual atmosphere and narrative quality of the image. For instance, adjusting lighting conditions can significantly impact the depth and brightness of the image, setting a mood can convey a specific emotional intent, and enhancing texture can create more visually engaging results.

By categorizing attributes into appearance, composition, and atmosphere, and

further refining them into specific parameters (e.g., color balance, style type, compositional ratio) using connection lines and slider functionality, this study provides a clear and efficient tool for prompt generation and image optimization. With this design, users can flexibly adjust single or multiple attributes to meet their specific objectives, enabling precise control and personalized expression in generated images.

6.4 Empty Frame and Connection

The Empty Frame module is an interactive image generation terminal node that implements attribute aggregation and AI generation fusion based on the Node class(see Fig. 6.2(C). Its core structure adopts a component-based design, consisting of two main parts: an SVG-based main input point located on the left side of the node, which handles attribute connections from other nodes through an event listening system, and a dynamically rendered hierarchical attribute panel that reuses the image-node attribute structure, containing three preset attribute groups: Appearance, Composition, and Atmosphere. The system monitors real-time connection states through a StateManager, triggering the Generate button activation logic when the number of valid connections reaches the preset threshold.

The connection function, the core mechanism for attribute transfer, implements dynamic connection lines based on SVG. Each connection point is designed as an interactive component, achieving color synchronization with sub-attributes through a CSS variable system. The connection system attributes three key characteristics: visual encoding based on attribute types, using specific color schemes to distinguish different attribute connections; adjustable weight control through connection line sliders enabling precise 0-100% control; and connection state visual feedback(see Fig. 6.7). When multiple image nodes connect the same attributes to a single empty frame, the system manages these attributes uniformly. It implements weight sharing and synchronized updates, with all related connection lines dynamically adjusting their visual representation accordingly(see Fig. 6.6).

PromptNavi implements two core mechanisms to optimize user experience in complex scenarios. The first is an attribute folding mechanism that employs a State Tree design(see Fig. 6.5), managing intra-group connections through recursive traversal. Users can choose to execute single-group or global attribute folding, with the system calculating the color blending of the main connection line through a weighted averaging algorithm. The second is a grouping management system based on map data structure, enabling efficient connection tracking and state updates. When users close a specific image node, the system automatically clears related connections and restores states, ensuring interface integrity. This design provides users with an intuitive and precise visual attributes editing system, transcending

the limitations of traditional text prompt input.

6.5 Attributes Interpolation

Behind the connection function lies a specially designed attribute interpolation mechanism, which enables intelligent attribute blending between images to help users optimize prompts more effectively. This study addresses several issues inherent in traditional text-based prompts, such as imprecise attribute descriptions, the need for repeated manual modifications, and inefficient exploration when using generative AI. Once a user creates an association between an image node and an empty frame via the connection system, the system processes and blends these attributes in multiple stages to generate new images.

The system adopts a dual-model attribute analysis framework. CLIP is used to deeply analyze the visual attributes of an image, extracting attribute vectors across three dimensions: appearance, composition, and atmosphere. A multilayer perceptron then processes these attributes to produce fine-grained sub-attribute classifications, such as color, style, and object under the Appearance category. A confidence mapping algorithm normalizes the model's output attribute weights into scores between 0 and 1, which are presented as intuitive percentage values in the interface. Meanwhile, a GPT-based deep semantic analysis module compensates for CLIP's limitations in contextual understanding by decomposing and reconstructing the original prompt, thus achieving a bidirectional complement between visual and semantic attributes.

An innovative layered-weight fusion algorithm underpins the attribute interpolation module. At the attribute level, the system first aggregates attributes, using a dynamic weight matrix to calculate the blend ratios of the same attributes across different images. For instance, when processing color attributes, the system constructs an attribute vector space, maps each image's color descriptions into this space, computes semantic distances between attributes via cosine similarity, and then performs intelligent interpolation based on user-defined weights. This approach ensures mathematical accuracy in attribute fusion and semantic coherence in the resulting visuals. The system also implements a cross-attribute compatibility check through a pre-trained attribute association matrix to evaluate and adjust the harmony of different attribute combinations.

PromptNavi employs a template-based dynamic prompt generator at the prompt synthesis stage. It first gathers all the keywords that users want to interpolate into the Empty Frame from the aforementioned connection system, then performs semantic reassembly of the weighted attributes for each category. These discrete attribute descriptions are transformed into coherent natural-language prompts using predefined linguistic templates. An attention-based importance ranking mech-

anism ensures that the most significant visual attributes receive sufficient representation in the generated prompts. The system can intelligently merge or differentiate closely related attribute descriptions by calculating semantic similarity, preventing redundancy or contradictions.

These processes create a cumulative attribute mechanism. Not only the newly generated images and their analyzed attributes are available for immediate reference and repeated use, but they also form a continuously growing experimental attribute database. Users can start from any previously generated image node to explore new attribute combinations, and the system automatically analyzes and combines these attributes to create images with novelty and potential. Every image generation thus provides valuable insights for subsequent creations, forming a continually evolving creative ecosystem.

The attribute interpolation and visual feedback approach emphasizes the accumulation of attribute knowledge, allowing users to precisely control and blend image attributes while fostering a sustainable platform for image-generation experiments. Users can conveniently reuse, combine, and optimize existing attribute combinations, transforming each generation attempt into a reusable creative experience. Compared to traditional text-based prompts, this visual method of attribute manipulation turns abstract textual descriptions into intuitive attribute connections, enabling users to understand better and control attribute combinations. Furthermore, adjustable weights offer precise control over attribute impact—which is difficult to achieve with traditional text-based methods. Most importantly, the system's attribute accumulation mechanism provides continual opportunities for optimization and iteration, converting each generation into a reusable experience and greatly enhancing both efficiency and quality in the creative process.

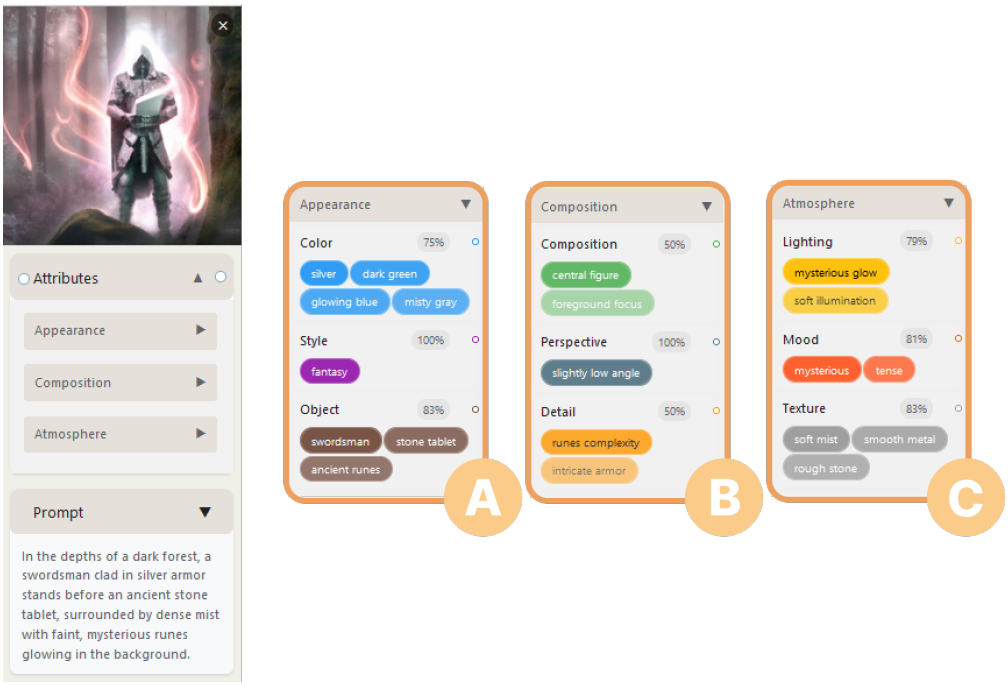


Figure 6.4: The visual attributes of the image node are organized into three main categories: Appearance (A) with Color, Style, and Object; Composition (B) with Composition, Perspective, and Detail; and Atmosphere (C) with Lighting, Mood, and Texture. Each attribute has a confidence score as a percentage indicating the strength of the attribute detection and an interactive connection point for linking attributes with other nodes.

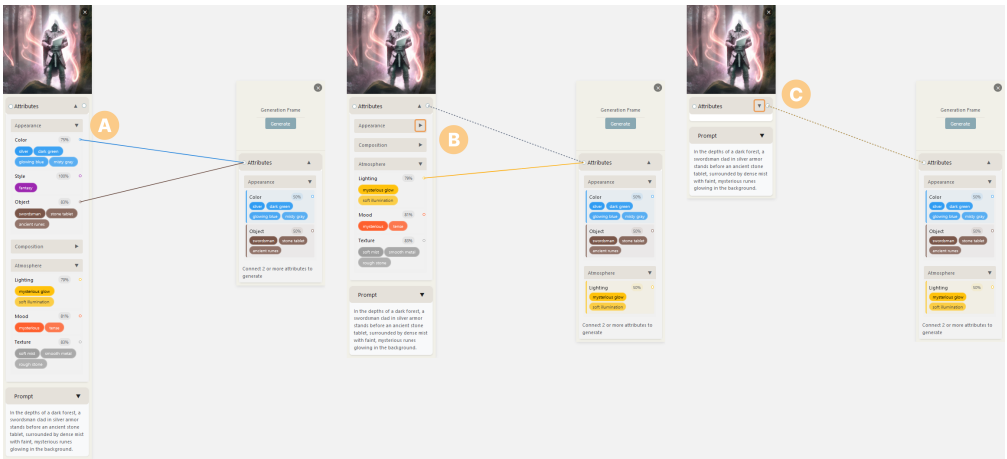


Figure 6.5: image node can link empty frame.

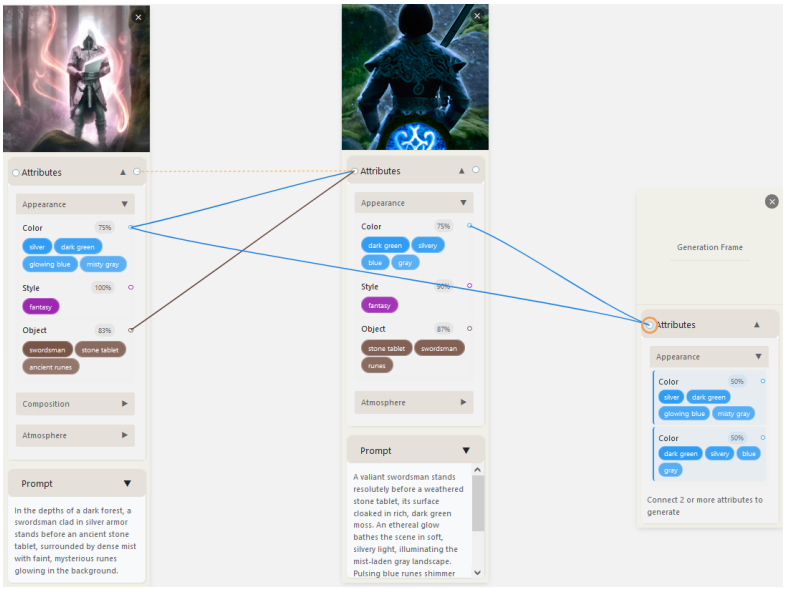


Figure 6.6: Different image nodes with the same attribute point can be connected to the same Empty Frame, which will then be treated as a single attribute.

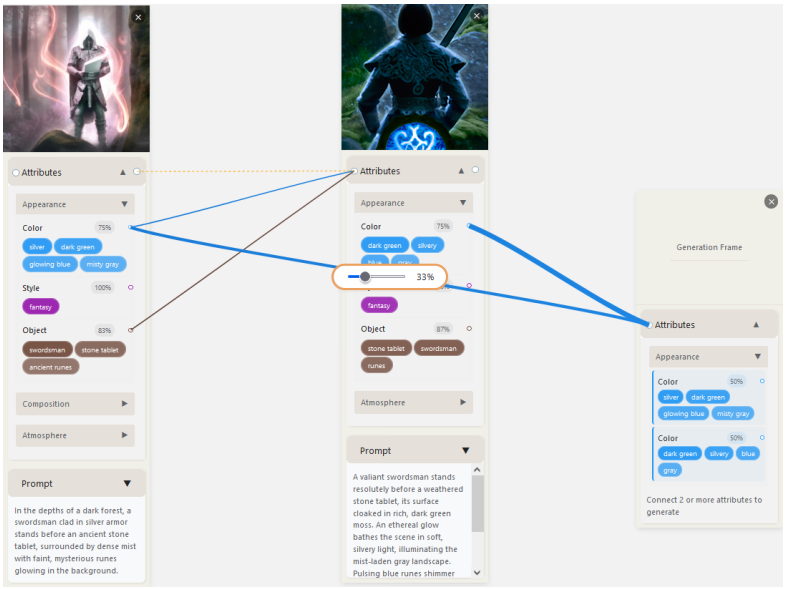


Figure 6.7: Double clicking on a connection line allows you to adjust its weight when the same attribute points from different image nodes are connected to the same Empty Frame. Once the weight is changed, it automatically updates all connections for that attribute to share the new weight, and variations in line thickness visually indicate this adjustment.

Chapter 7

User Study

This chapter presents the user study conducted to evaluate PromptNavi's usability and effectiveness, comparing it with a baseline system through tasks and feedback.

This study evaluates PromptNavi's innovation, its iterative optimization and generation quality capabilities, and its actual performance in assisting users with generation tasks. To achieve this, a comparative evaluation was conducted, focusing on the proposed system's performance differences in multi-round image generation tasks. The study compares usage efficiency, image quality, and overall user experience to determine whether PromptNavi's design can significantly improve users' efficiency and effectiveness in text-to-image generation tasks.

7.1 Baseline Approach

For this study, only the text-to-image (txt2img) function in Stable Diffusion WebUI was referenced, as it generates images solely based on textual descriptions without requiring additional input modalities such as image references or sketches. This approach was chosen to avoid bias and minimize subjective intervention in image generation. Only the most essential interface elements—namely the prompt input area, image display area, and generate button—were retained, and this configuration was used as the baseline system. The baseline shares the exact implementation approach as the proposed system, employing HTML and JavaScript for the front end and Flask for the back end.

An online pre-trained model (DALL-E) was used to ensure consistent image generation and highlight differences in prompts. PromptNavi is also used with identical settings. A within-subject design was adopted to compare user experiences between PromptNavi and the baseline system. Half of the participants started with PromptNavi to balance usage order, while the other half began with the baseline.

7.2 Procedure

This study consists of two phases: an exploratory phase and a specific creative task. In the exploratory phase, to help participants explore creative possibilities more efficiently, they are encouraged to select a thematic style (e.g., "Fantasy Forest" or "Future City") as their starting point. Alternatively, participants may propose their creative theme if desired. This initial prompt remains consistent across both systems and is the starting point for image generation and iteration. Participants can freely expand or modify the theme in subsequent iterations, allowing for natural creative exploration. After completing the exploratory phase, participants proceed to the specific task phase where they are asked to create an image with defined requirements: "a futuristic city street at dusk" with specific required elements.

Participants received a brief introduction to the study and a walkthrough of each system, during which they operated the systems under the authors' guidance. Each system was introduced with a 5-minute explanation. The generated image and its associated prompt were saved. Then, participants tested both systems independently and completed corresponding questionnaires. After completing the comparative evaluation, participants engaged in a semi-structured interview. Throughout the study, participants were encouraged to verbalize their thoughts and ask any questions they had.

Exploratory Flow

- **Step 1** Participants input a basic prompt (e.g., "Fantasy Forest, sunlight, magical elements") into the current system to generate the first image and corresponding prompt.
- **Step 2** Participants select their favorite attributes from the image and prompt, such as specific visual attributes or descriptive keywords.
- **Step 3** Participants adjust the prompt to generate a new image using the selected attributes. During this process, they can freely choose elements from all previously generated images to refine the prompt and produce better results.
- **Step 4** The process is repeated until five images are generated. Each step records all images, corresponding prompts, and iterative adjustments.
- **Step 5** After completing the tasks with the first system, participants take a short break to reduce fatigue and mitigate potential learning effects.

- **Step 6** Participants switch to the other system (PromptNavi or Baseline) and repeat the same procedure from the initial generation to iterative optimization.

Requirements

No strict limit was imposed on the final number of prompts or generated images. However, participants were instructed to carry out five rounds of iteration:

In each round, participants could modify the prompt any number of times and generate as many images as they wished until they felt the result was "sufficiently satisfactory."

Then, from all the images produced in that round, they would choose one that they considered the most ideal, along with its corresponding prompt.

The next round of iteration would be based on the chosen image and prompt from the previous round, continuing to refine, expand, or explore new creative directions.

In this way, each participant went through five rounds of iteration. However, there was no strict limitation on the number of times they could modify the prompt or the number of candidate images they could generate in each round. The only requirement was that they keep one final "most satisfactory" image and its prompt from each round.

Specific Creative Task

Participants were eventually asked to create a final image depicting a city at dusk with a rich humanistic atmosphere to provide a common baseline for evaluating the system's performance. While this scene should convey a sense of everyday human life, participants have ample freedom in representing modern architecture, a warm dusk atmosphere, and bustling street elements such as pedestrians or vehicles. This specific request serves as a controlled reference point, complementing the insights gained from the exploratory iterations and offering a unified benchmark for assessing the system's creative output. After the work was submitted, external individuals who had not participated in the experiment were invited to evaluate and score the pieces. Their assessments will draw on multiple dimensions, including creativity, atmosphere, coherence, and overall impression, providing a more objective and diverse perspective on the quality of the creative outcomes and the system's performance.

Semi-Structured Interview

After completing two study phases with both systems, an in-depth interview lasted approximately 10–20 minutes. The discussion focused on the following three usage

scenarios:

- For Scenario A: Repetitive Iteration While Maintaining Style Consistency

  Participants were asked how effectively the system supported 'batch modifications,' retained key style elements, and maintained efficiency in this process.

- For Scenario B: Refining a Vague Concept Step by Step

  Participants were asked about the system's ability to help them express initial fuzzy ideas, add details incrementally, and provide layered control throughout the creative process.

- For Scenario C: Open-Ended Exploration and Inspiration

  Participants were asked whether the system was conducive to rapid experimentation, flexible style shifts, and inspiration-seeking in the absence of a specific goal.

Interviews also allowed participants to share any challenges encountered, opinions on system features, and suggestions for improvement, providing deeper qualitative insights.

7.3 Participants

A total of 16 volunteers were recruited from within the university to participate in this study, consisting of 12 male and 4 female participants. All participants had varying degrees of experience with Generative AI, and their ages ranged from 22 to 32. In addition, all participants were graduate students with an HCI background, ensuring they had a fundamental understanding of the research tasks.

Before the experiment officially started, each participant received a brief overview of the procedures and training on using relevant tools, ensuring they were familiar with the experimental process and the basic functionality of the tools involved. This study strictly adheres to research ethics guidelines and privacy protection requirements, and all collected data will be used solely for research analysis.

7.4 Setup

The study was conducted in a quiet laboratory environment. A Microsoft Windows laptop (Intel i7-11800H, NVIDIA GeForce RTX 3070) capable of running both the proposed system and the baseline simultaneously was placed on a desk, accompanied by a large external monitor, a mouse, and a keyboard. Participants

interacted with the proposed system and the baseline according to the experimental procedure(see Fig. 10.1).

7.5 Data Collection

Screen recordings were captured during each task, and participants' explanations of their behaviors were documented. The time participants took to complete the tasks and the time spent on each iteration during the generation process were recorded for usage efficiency. For image quality, a subjective rating system designed for this experiment was employed to compare the quality of prompts and generated images. Participants' feedback was collected under each condition. All ratings used a 5-point Likert scale (1 – Strongly Disagree, 5 – Strongly Agree) [59]. For overall system usability, the System Usability Scale (SUS) [60] (see Table 10.2) was employed to assess participants' perceptions of the system's ease of use and user satisfaction, while the NASA-Task Load Index (NASA-TLX) [61] (see Table 10.1) was used to evaluate cognitive load during task completion.

The specific subjective rating criteria are as follows: specific subject rating questionnaires are Table 10.3:

- *Visual Quality:* Whether the image's content is rich and visually appealing.
- *Element Integration:* Whether the key elements extracted from previous prompts or images are effectively integrated into the final output, and whether the result meets expectations.
- *Style Consistency:* Whether the image style aligns with the original theme of the generation task.
- *Optimization Effectiveness:* Whether the final image reflects the improvements made during the iterative process.
- *Exploration Support:* Whether the system supports flexible adjustments and exploration to help users generate satisfactory images.

Chapter 8

Results

This chapter presents the user study results, highlighting PromptNavi's advantages in usability, creativity, and efficiency compared to the baseline system.

8.1 Significant Improvements in PromptNavi's Usability and User Favorability

A comprehensive evaluation of PromptNavi has demonstrated notable enhancements across multiple dimensions compared to the baseline system. Assessments using the System Usability Scale (SUS) indicate that PromptNavi achieved an overall usability score averaging 80.31 (median 82.50), while the baseline system scored 57.50 (median 56.25). This difference (Cohen's d = 1.931) signifies a highly significant practical improvement in usability, reaching what is traditionally classified as a "large effect size."

From the box plot(see Fig. 8.1), it is verified that PromptNavi's scores are higher on average and more tightly clustered, indicating that most participants share a relatively consistent, positive perception of its usability. In contrast, while the baseline system does feature a few higher outliers, its overall distribution leans toward lower scores with a greater spread, implying that user experiences vary considerably. However, given that the baseline is both a traditional and widely used system, the large gap in scores could partly stem from participants' heightened favorability toward PromptNavi, and the possibility that novelty effects may have lowered the baseline scores cannot be entirely ruled out.

The per-item distribution chart provides a more detailed view of participants' attitudes toward specific usability dimensions (see Fig. 8.7), supplementing the broader usability analysis. For instance, regarding the statement "I thought the system was easy to use," over 80% of participants selected "Agree" or "Strongly Agree," indicating that most found PromptNavi's core interactions and workflow

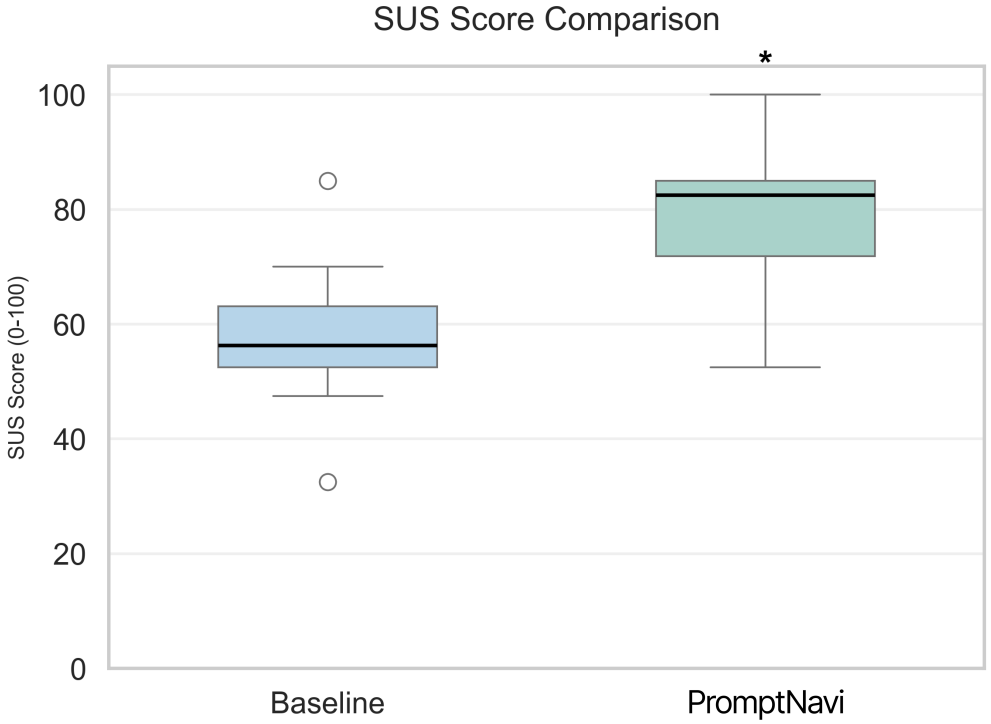


Figure 8.1: Results of SUS total scores: Items marked with an star(*) indicate statistical significance (p<0.05). The baseline (n=16) had a mean score of 57.50 with a median of 56.25. PromptNAVI (n=16) had a mean score of 80.31 with a median of 82.50. The effect size between the two groups was Cohen's d = 1.931, indicating a large practical significance in the difference.

to be intuitive and low in complexity. Meanwhile, over 90% expressed or strongly agreed with "I found the various functions in this system were well integrated," suggesting that users perceived the system's modules as working cohesively together, reducing extra burdens when switching between different features. This observation is consistent with Hypothesis (H2), which posits that the system's prompt interpolation functionality helps reduce the repetitive task of crafting prompts.

Regarding the item "I think I would need a technical person's support to use this system," more than half of the participants chose "Strongly Disagree" or "Disagree," with only a handful indicating a need for advanced external assistance. According to participant feedback, this is mainly because the node-and-connection design mirrors familiar interaction paradigms in many existing products. Users can intuitively attach attributes to a new target (empty frame) via a connecting line, granting the target that attributes straightforwardly and logically. This finding aligns with the widespread agreement on the statement, 'I would imag-

ine that most people would learn to use this system very quickly,' indicating that the majority of participants perceived PromptNavi's learning curve as relatively gentle. This result supports Hypothesis (H4), which contends that the system offers superior practicality and ease of use compared to traditional generative AI interfaces. Additionally, for negative statements such as "I thought there was too much inconsistency in this system," most participants responded with "Disagree" or "Strongly Disagree," further confirming the high ratings regarding the system's consistency.

8.2 Substantial Decreases in Workload and Gains in Performance

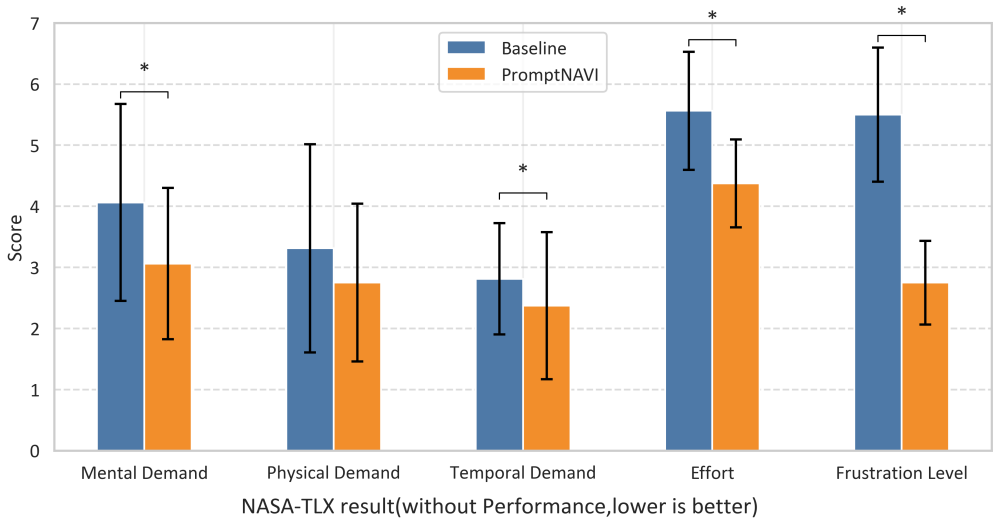


Figure 8.2: Results of task load ratings(without Performance,lower is better): Items marked with an star(*) indicate statistical significance (p<0.05).

The NASA Task Load Index (NASA-TLX) shows that PromptNavi outperforms the baseline system in the five negatively scored dimensions—Mental Demand, Physical Demand, Temporal Demand, Effort, and Frustration Level—and also achieves a significantly higher Performance score(see Fig. 8.2 Fig. 8.3). Interviews reveal that PromptNavi's node-and-connection interface reduces repetitive prompt entry and makes grasping the relationship between prompts and outputs easier. The notable decrease in Mental Demand and Effort aligns with H1 and H2, as users can visually compare and reuse prompts in one workspace instead of frequently switching between text fields or scripts. Some participants also reported that this "what you see is what you get" approach maintained their focus

Metric	Baseline	PromptNAVI	Effect size
Mental Demand	4.06	3.06 *	0.87
Physical Demand	3.31	2.75 †	0.49
Temporal Demand	2.81	2.38 †	0.38
Effort	5.56	4.38 *	1.03
Frustration Level	5.50	2.75 *	2.42

* indicates <0.001; † indicates <0.05

Table 8.1: NASA-TLX Scores Comparison(without Performance, lower is better

Metric	Baseline	PromptNAVI	Effect size
Performance	3.19 ± 1.47	6.38 ± 0.78 *	2.73

* indicates significant difference ((p<0.001))

Table 8.2: Performance Score Comparison(higher is better)

during iterative adjustments while alleviating the anxiety of blind trial-and-error. In contrast, baseline users often re-generated the same prompt and endured long "idle" waits. In contrast, in PromptNavi, they could keep refining ideas through other node-based interactions while generation takes place.

Physical and Temporal Demands were improved. PromptNavi's drag-and-drop design reduces constant keyboard-mouse switching, and with fewer repetitive text inputs and attribute interpolation, users can more readily explore different styles in a limited time. Interviews indicated that baselines often forced participants to switch between multiple windows to document and compare prompts. In contrast, PromptNavi centralizes these tasks via nodes and connections, cutting down on interface toggling. Its advantage in frustration level further underscores the positive role of visualizing prompts: rather than comb through lengthy text prompts to locate errors, users can directly edit attributes and connections to isolate issues—supporting H1, H2, and H3 by minimizing ineffective attempts, reducing prompt-writing burdens, and enabling satisfying results. In addition, users can simultaneously employ node-based and text-based approaches if desired.

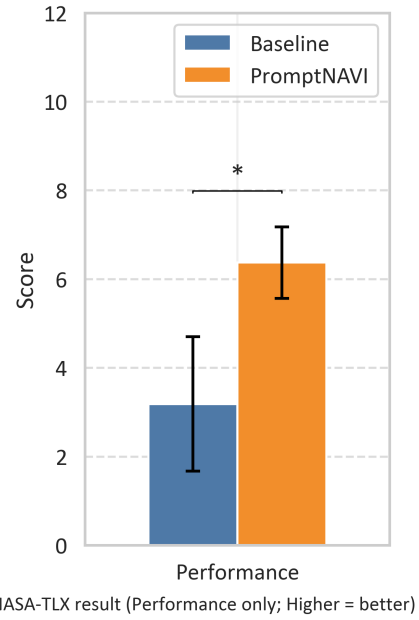


Figure 8.3: Results of task load ratings(Performance, higher is better): Items marked with an star(*) indicate statistical significance (p<0.05).

Metric	Baseline	PromptNAVI	Effect size
System Usability Scale	53.75 ± 15.91	80.63 ± 12.37 *	1.89

* indicates <0.001

Table 8.3: Results of SUS evaluation.

8.3 Elevated Satisfaction and Creative Control

In the Subject Rating assessment, PromptNavi scored significantly higher than the baseline system across all 10 evaluated metrics (see Fig. 8.9, where higher scores indicate greater satisfaction or approval). Participants gave particularly positive feedback on “visual appeal” (Item 1) and “content richness” (Item 2), which aligns with interviews mentioning the system’s detailed and refined outputs. Users could freely add or combine key elements by supporting visual attribute management and quick prompt reuse, resulting in images that better matched their aesthetic and structural expectations (supporting H5).

In “key element integration” (Item 3) and “consistency with expectations” (Item 4), PromptNavi’s advantage became more apparent. Many users noted

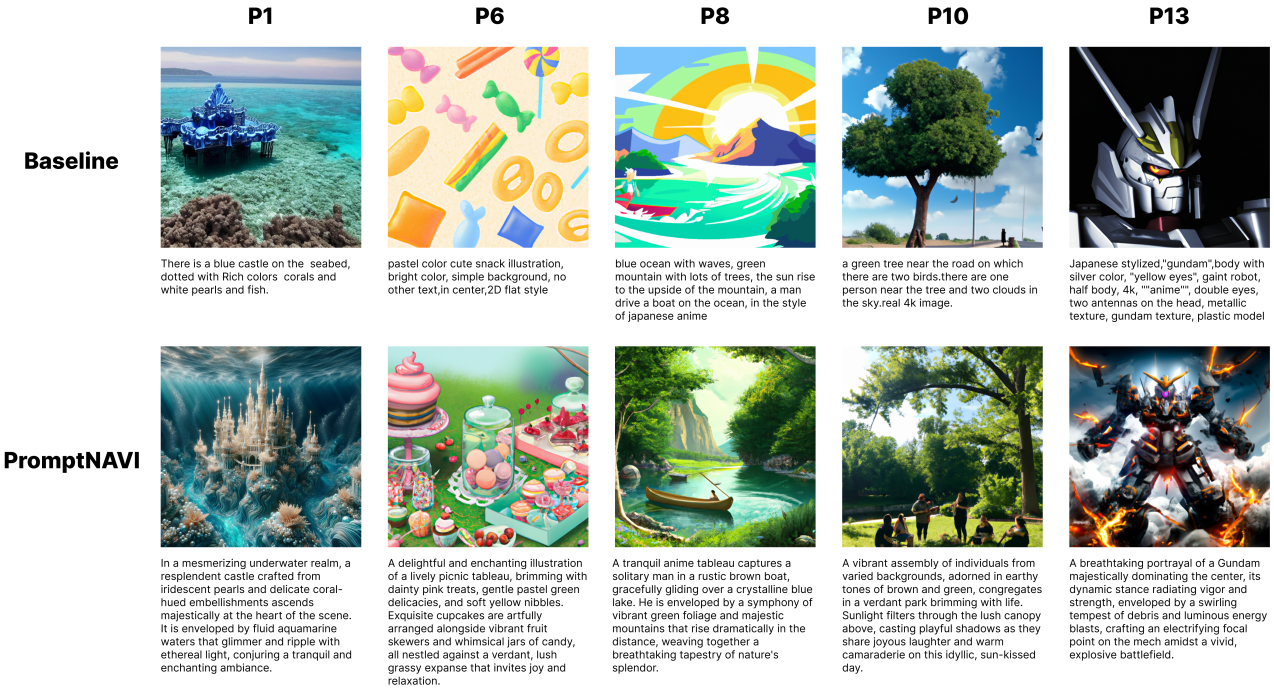


Figure 8.4: Example prompt with images from the result of explore study by participants using PromptNavi and Baseline system.

that its node-and-connection interface allowed them to visually track and adjust previously tested prompts without sifting through raw text. This "what you see is what you get" approach boosted efficiency (H1, H2) and ensured that core styles and elements remained consistent with the initial concept.

PromptNavi also outperformed the baseline in style consistency (Items 5, 6) and steady improvement across multiple iterations (Items 7, 8). Participants with more substantial design needs highlighted that the visual linking of attributes helped maintain coherent styles when generating related images; if they wanted to add new ideas or modify existing elements, they could confirm and execute changes quickly without undermining the overall style. These findings align with the reduced workload in NASA-TLX and support H3 (achieving more satisfying outputs) and H5 (fulfilling users' artistic goals).

Finally, PromptNavi excelled in system flexibility (Item 9) and compatibility with new ideas (Item 10). Several participants reported that adding new creative elements in the baseline system required rewriting large portions of text prompts, which often conflicted with existing style instructions. In contrast, PromptNavi lets them add, move, or remove nodes in a visual interface or use prompt inter-

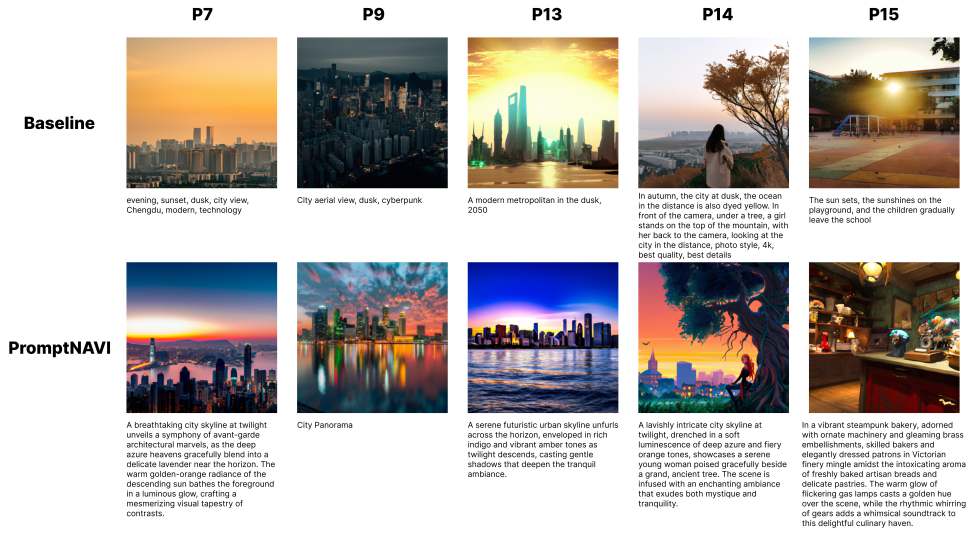


Figure 8.5: Example prompt with images from the result of Specific Creative Task by participants using PromptNAVI and Baseline system.

polation to incorporate new styles or attributes smoothly. This flexibility reduced repetitive prompt drafting and made the iterative design more convenient (supporting H2, H4), opening up broader creative possibilities.

Combined with usability (SUS) and workload (NASA-TLX) findings, these results confirm that PromptNavi lowers users' cognitive and operational burdens and significantly enhances visual quality, style consistency, iterative improvements, and flexibility. In other words, in addition to minimizing redundant attempts and inputs (H1, H2), PromptNavi helps produce more satisfying outputs (H3), provides a more intuitive and adaptive workflow (H4), and effectively supports users' creative visions (H5). These comprehensive improvements reinforce the effectiveness of node-based visualization and prompt interpolation.

8.4 High Efficiency with Potential Diminished Awareness of Prompt Iteration

During the exploratory phase, two main aspects were examined: image and prompt richness, along with the completion time across multiple iterations. In terms of image richness, many participants indicated that PromptNavi's node-and-connection interface allowed them to retain and recombine appealing attributes from previously generated images more efficiently. This often resulted in visually richer and more detailed outcomes than the baseline system, where participants had to re-

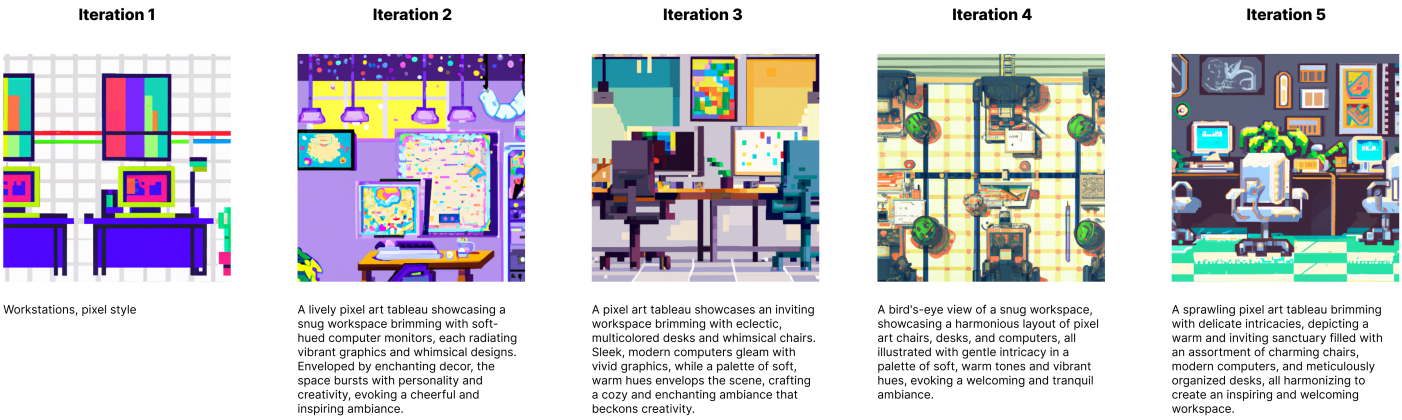


Figure 8.6: An example result of iterative step a prompt explore with an image.

peatedly modify text prompts and generate new images from scratch. Combined with earlier Subject Rating findings (e.g., 'visual appeal' and 'content richness'), this observation confirms that PromptNavi enables users to achieve higher-quality images with fewer text-based revisions (see Fig. 8.9).

Regarding prompt richness, the two systems showed different iteration patterns(see Fig. 8.5). In the baseline system, iterative changes were reflected in the evolving text prompts, enabling users—and observers—to trace how specific keywords or descriptions became more refined over time. In PromptNavi, however, participants often relied more on visual attributes and less on editing text once they became comfortable with the interface. As a result, the final prompt did not always capture the entire iteration history(see Fig. 8.6). While this approach can reduce the cognitive burden of constant text editing, it may also limit the ability to review the precise textual iteration of prompts.

Regarding completion times, PromptNavi typically reduced the time needed for the first few iterations, likely because users did not have to rewrite prompts extensively. As iterations progressed, the interface accumulated more nodes and attributes, providing greater creative flexibility at the expense of slightly longer interaction times. In contrast, baseline users often spent significant time making repeated minor changes or attempting dozens of random generations, especially in a single iteration, due to the lack of a visual framework. Consequently, although the baseline system sometimes resulted in large blocks of text that revealed clear iterative traces, these iterations often required more trial-and-error and led to highly variable performance.

Overall, PromptNavi appears to help users create richer images with fewer direct text edits, thanks to its node-and-connection system that stores and reuses attributes across iterations. However, this visual approach can downplay the it-

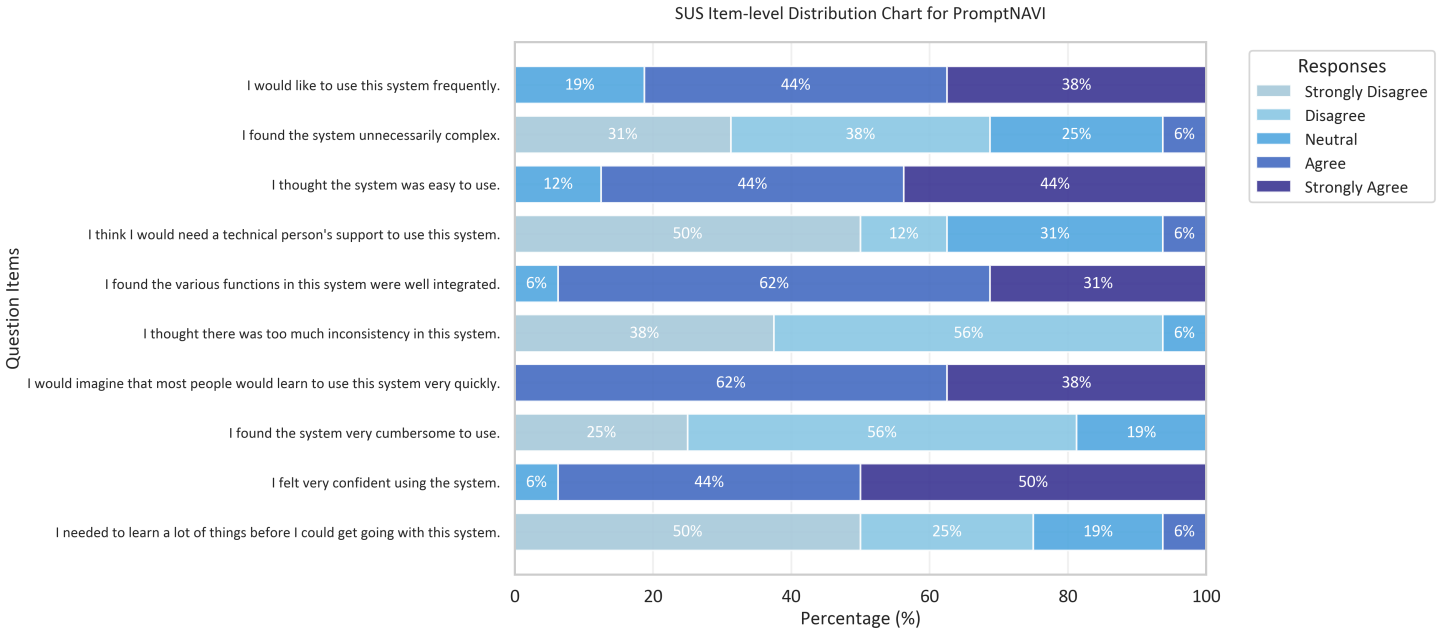


Figure 8.7: SUS Responses Distribution Chart for PromptNAVI.

eration of text prompts. While users might spend a bit more time per iteration once many attributes accumulate, they also benefit from greater flexibility and control during the creative process. Balancing the system's strengths in visual attribute management with the benefits of textual traceability remains a key design consideration in the future.

8.5 Superior Performance in Third-Party Ratings

In this specific creative task, participants were instructed to focus on "a city at dusk," emphasizing a warm twilight atmosphere, a humanistic vibe, and diverse street elements. To evaluate the final outputs, 20 graduate students who had not participated in the earlier experiments were recruited to assess the results based on four dimensions: creativity, atmosphere, coherence, and overall impression. (see Fig. 8.10 and Table.8.5) Comparing the average and median values shows that PromptNavi achieved significantly higher scores than the Baseline system in all four dimensions, with statistical significance ($p<0.01$).

Based on the score analysis, the images produced by PromptNavi demonstrated more consistent coordination among lighting, color, and human elements, contributing to a layered warmth in the dusk setting. Regarding Creativity, Prompt-

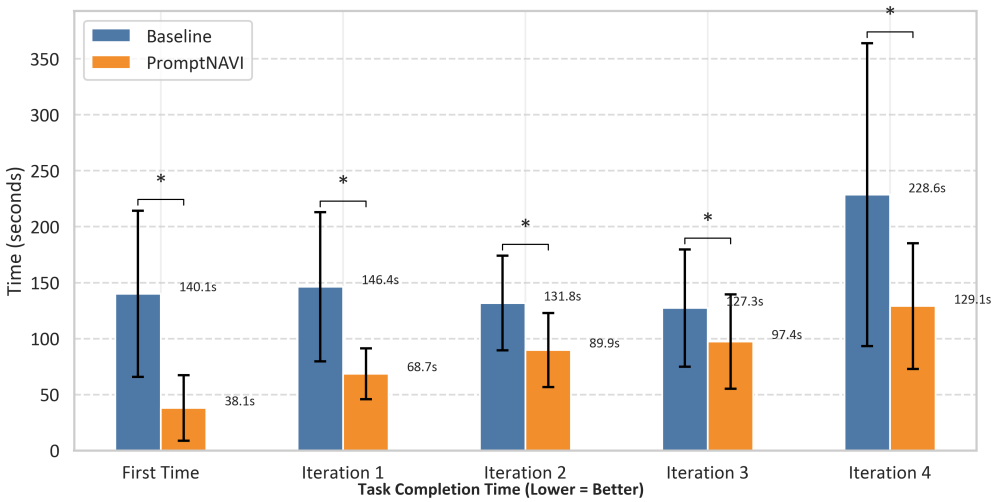


Figure 8.8: Results of iterate step time in exploratory phase: Items marked with a star(*) indicate statistical significance (p<0.05).

Navi's scores were more concentrated in the higher range, indicating it supported more decadent combinations of urban architecture, street layouts, and crowd activities—common scenes included pedestrians strolling, vehicles in motion, and small shops. The baseline system occasionally produced highly creative results as well, but on the whole, these varied more widely, and many works appeared somewhat monotone in capturing the theme. In terms of Atmosphere, PromptNavi excelled by leveraging its node-and-connection interface to amplify the sunset lighting layers and the street environment—for instance, by highlighting warm gradients in the sky and seamlessly integrating crowds into the urban background—thus more closely reflecting viewers' expectations for a twilight cityscape.

PromptNavi again scored higher for coherence, with a narrower interquartile range, suggesting that most works delivered cohesive building scales, pedestrian proportions, and light directions with fewer missing or incongruent details. In contrast, Baseline works sometimes omitted key street elements or exhibited abrupt color transitions, compromising overall visual continuity. Finally, in Overall Impression—an aggregate measure of aesthetic appeal, completeness, and thematic alignment—PromptNavi led by a notable margin. Its 75th percentile significantly surpassed the Baseline's, indicating that most PromptNavi-generated images offered a more striking and rich human atmosphere twilight city ambiance.

These four rating dimensions suggest that PromptNavi more readily maintains a warm color scheme, richer urban elements, and a pronounced human touch, leading to consistently higher-quality final results. This outcome aligns with previous findings from the exploratory phase. By harnessing visual attribute management

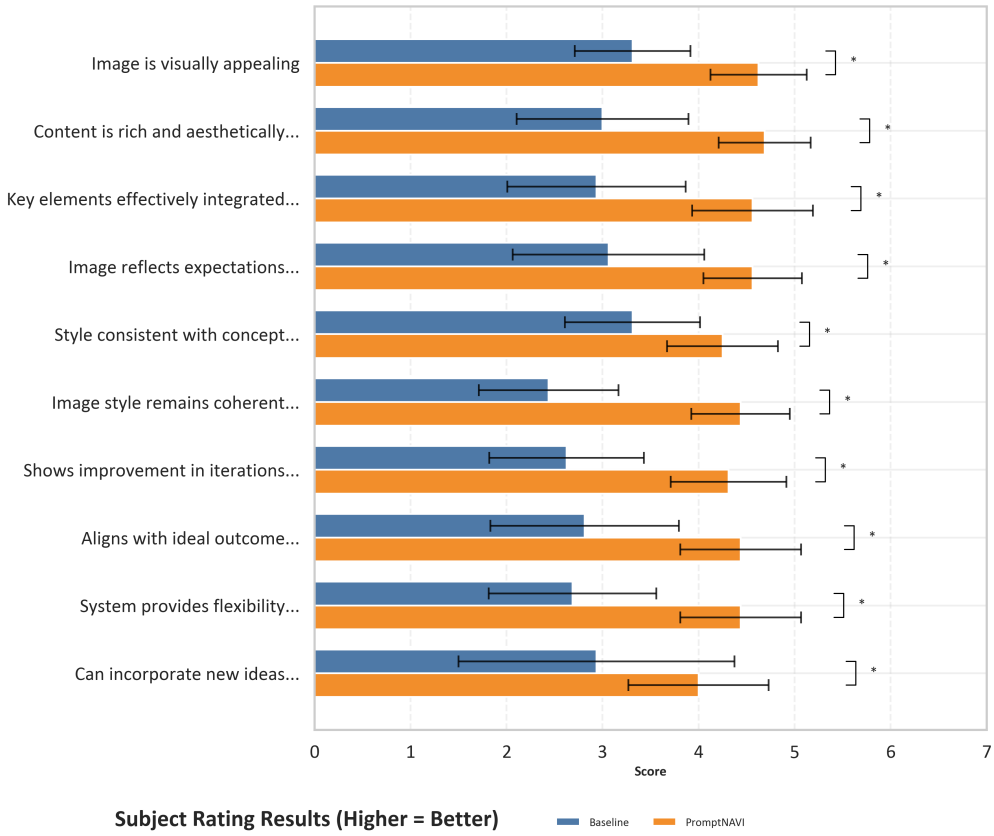


Figure 8.9: Results of subject ratings(Table 10.3): Items marked with an star(*) indicate statistical significance (p<0.05).

and prompt interpolation, users can quickly preserve and refine promising details at each iteration, reducing the trial-and-error burden. However, it is worth noting that participants' varying interpretations of the "city at dusk" theme also influenced the final content, with some emphasizing a commercial feel and others focusing on human-scale storytelling, potentially affecting evaluators' priorities. These data confirm that PromptNavi delivers both strong usability and high production quality for a visually guided, human-centered creative prompt, laying a solid foundation for further applications in scene-based design.

Metric	Baseline	PromptNAVI	Effect size
Image is visually appealing	3.31 ± 0.79	4.62 ± 0.50 *	2.37
Content is rich and aesthetic	3.00 ± 0.89	4.69 ± 0.48 *	2.35
Key elements effectively integrated	2.94 ± 0.93	4.56 ± 0.63 *	2.05
Image reflects expectations	3.06 ± 0.85	4.56 ± 0.73 *	1.89
Style consistent with concept	3.31 ± 0.70	4.25 ± 0.58 *	1.46
Image style remains coherent	2.44 ± 0.63	4.44 ± 0.63 *	3.18
Shows improvement in iteration	2.62 ± 0.81	4.31 ± 0.60 *	2.37
Aligns with ideal outcome	2.81 ± 0.91	4.44 ± 0.73 *	1.97
System provides flexibility	2.69 ± 0.79	4.44 ± 0.73 *	2.30
Can incorporate new ideas	2.94 ± 1.18	4.00 ± 1.10 †	0.93

* indicates <0.001; † indicates <0.05

Table 8.4: Subjective rating comparison

Metric	Baseline	PromptNavi	Effect size
Creativity	2.79 ± 1.22	3.42 ± 1.10 *	0.54
Atmosphere	2.84 ± 1.19	3.44 ± 1.15 *	0.51
Coherence	2.84 ± 1.10	3.30 ± 1.11 *	0.42
Overall	2.80 ± 1.06	3.35 ± 1.04 *	0.52

* indicates significant difference (p<0.001)

Table 8.5: Table of Third-party evaluation of iteration images

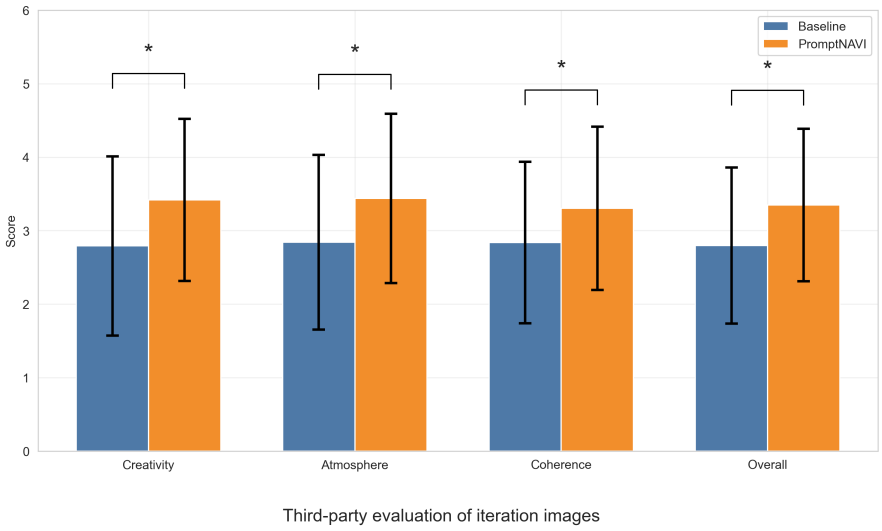


Figure 8.10: Third-party evaluation of iteration images: Items marked with an star(*) indicate statistical significance (p<0.05).

Chapter 9

Discussion

After the study, semi-structured interviews were conducted with participants to gather feedback on batch editing, repeated iterations, core style retention, and suggestions for improvement. The following discussions summarize the responses.

9.1 Part One: Batch Editing and User Experience

14 participants acknowledged the convenience and visual advantages of the batch editing feature. Using graphical connections and attribute dragging, they no longer need to manually input or remember complex Prompt phrases during each iteration; instead, they can more intuitively understand and control each attribute. This design makes the process more straightforward and reduces the mental burden of repeated modifications. 2 participants mentioned that automatic summarization and Prompt generation can lower the amount of trial-and-error, which is particularly helpful for users who are less familiar with the specifics of prompt writing.

Nevertheless, some notable challenges emerged in practical operation. 3 participants reported that the drag-and-drop process is susceptible to click position, with connecting nodes sometimes being too small, leading to difficulties in selecting or inadvertently misconnecting. In addition, specific attributes do not allow enough editing flexibility, making it impossible to delete unnecessary or fine-tune different tags under the same attribute. For first-time users, facing many attributes with non-intuitive names can increase the difficulty of manipulation and lead to confusion during batch editing. While the batch editing feature excels in providing a "visually intuitive" approach, it still requires a better balance between user-friendly interfaces and more detailed operations.

9.2 Part Two: Repeated Iterations and Core Style Retention

All participants offered positive feedback regarding style retention across multiple iterations. They generally observed that once the system moves through several iterations, the core style users wish to preserve often remains consistent, particularly in aspects such as the theme, structure, or key visual attributes, where the system seems to "edge closer to the desired goal" step by step. If users want more specific changes, they need only adjust the weight of the relevant attributes to see a noticeable difference in the next iteration.

However, 2 participants noted occasional unintended shifts in details, especially with color or minor decorative attributes, which might fade or change over multiple iterations. In such cases, many opted to increase the desired attributes' weight or revisit previously successful keywords to "pull the style back." Some preferred to create a brand-new image node when encountering significant deviations to avoid wasting time repeatedly correcting mistakes in the same environment. These observations suggest that while the system has a specific capacity for repairing minor deviations in the subsequent round of iteration, users also need to master proper adjustment techniques or plan for necessary rollbacks and restarts.

9.3 Part Three: Process Efficiency, Suggestions for Improvement, and Iteration Insights

Participants generally agreed that a visual approach is faster and more intuitive than typing prompts, particularly for rapid prototyping or frequent edits. Some suggested a "regenerate" option to avoid repeatedly reconnecting attributes, thereby keeping previous results while only adding or modifying new ones. Others emphasized the need for a more refined rollback mechanism—such as undoing only the last step or deleting specific attributes—so users aren't forced to restart entire workflows.

When asked how they might seek new inspiration when lacking a clear direction for iteration, 80% participants indicated they would review the images already generated, determine which outcomes matched expectations, and pinpoint which attributes needed to be strengthened or diminished. Others rely more heavily on the system's automatic expansion features or attribute suggestions, using this guidance to spark fresh ideas. Once they have a clearer vision of the desired style, they often return to manual intervention and weight adjustments. This indicates that the process is a balanced interplay between machine guidance and human expertise. Maintaining high efficiency in batch editing and repeated iterations requires

users to integrate existing results, automated system prompts, and experience, ensuring a smooth and productive creative workflow.

9.4 Key Findings

4 key conclusions can be drawn based on the discussion in the three sections above.

- 1. Employing a visual linking mechanism and automatic prompt summarization for batch editing significantly lowers the barrier to entry while improving operational efficiency in iterative scenarios.
- 2. The system proves relatively stable in retaining the core style, especially in multiple iterations where themes and principal attributes remain preserved mainly; however, more flexible weight adjustments or rollback mechanisms may be required to manage color details or other nuanced aspects.
- 3. Accidental loss of style is typically resolved by shifting attribute weights or creating a fresh environment, and how well users handle these fixes directly affects the outcome. In addition, the drag-and-drop interface and attribute editing functionality still need refinement—mainly when numerous attributes with obscure names leave users unsure of their roles.
- 4. Without a clear iterative direction, participants often rely on existing outputs or automated expansions for inspiration and refine their ideas through manual modifications—underscoring the importance of dynamic interaction between the system and the user. Nevertheless, the lack of more profound or granular editing still leaves users with random fine-tuning and filtering. Future improvements should thus focus on offering more precise editing tools and assistance.

Chapter 10

Conclusion

10.1 Conclusion and Futurework

This work presents PromptNavi as an interactive system designed to address the challenge of prompt optimization in text-to-image models. PromptNavi transforms the traditionally trial-and-error-heavy prompt editing process into a more intuitive, visually guided workflow by incorporating a node-based visual interface, dynamic connection mechanisms, and an attribute interpolation technique. Both exploratory experiments and a comparative study with a specific creative task were conducted to comprehensively evaluate PromptNavi's performance in multi-iteration editing, batch modifications, and final output quality. Compared with the baseline system, the results show that PromptNavi significantly improves user satisfaction, generation efficiency, and iterative image quality. On the one hand, questionnaire, and interview data confirm its effectiveness in reducing users' cognitive load; on the other, its interpolation and visual management features provide deeper insight into how prompts influence final outputs, enabling novice and experienced users to achieve their creative goals more flexibly.

Despite these promising findings, there remain several directions for further exploration. First, enhancing the traceability of prompt changes is vital for optimizing large-scale or complex designs, as it would allow users to systematically revisit and reuse successful elements from each iteration, thereby boosting creative efficiency. Second, introducing fine-grained attribute control and a more flexible interface layout—such as customizable attribute content, more robust line insertion/deletion, and options for regenerating or extending Image Node—will help accommodate high-complexity or professional-level tasks. As these enhancements are gradually implemented, PromptNavi is poised to deliver broader and more robust support in various creative scenarios, offering novel insights and practical strategies for interactive design and prompt optimization in text-to-image genera-

tion.

10.2 Limitations

Despite the system's clear benefits in lowering creative barriers, streamlining batch editing, and preserving overall style, several limitations warrant attention.

First and most importantly, the traceability of prompt iteration is somewhat lacking. Because users mostly rely on visual operations rather than frequent text edits, the final prompt does not always capture each iteration's key changes. In traditional text-based systems, every revision is recorded as a new version, offering an advantage for users needing to review or reuse specific prompt variations. Although PromptNavi's streamlined process reduces text input, it obscures detailed prompt changes, potentially causing users to ignore the prompt layer entirely. This may become a future trend for some workflows but could limit those who rely on thorough text records.

Second, connection functions can become cumbersome when dealing with many attributes or highly complex themes. Some participants reported difficulties when nodes were too small or densely arranged, making accidental clicks or mislinks more likely. Also, specific attributes have abstract names that new users find hard to interpret, reducing operational efficiency.

Lastly, although the system maintains an overall style across multiple iterations, fine details (e.g., colors or decorative elements) sometimes drift over time. Participants often adjust attribute weights or revert to previous keywords to recapture lost details; if errors become too pronounced, they create a new image node rather than repeatedly fixing the same workspace. While PromptNavi can self-correct within a small range, exact aesthetic consistency relies heavily on user skill.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Haoran Xie, for his invaluable guidance and unwavering support throughout my studies and research. Prof. Xie's profound knowledge and insightful perspectives have significantly influenced the direction and quality of my work. At the same time, his emphasis on rigor and clarity led me to a deeper understanding of academic excellence and thorough research. His wide-ranging ideas and innovative insights continually pushed me to pursue higher standards, helping me refine my academic thinking and exploration.

In addition, I sincerely appreciate my former collaborating professor—Chia-Ming Chang, currently an assistant professor at the National Taiwan University of Arts, and my former collaborator—Xinyue Gui, now a PhD candidate at the University of Tokyo. In the early stages of my academic journey, they provided abundant research support and invaluable guidance, particularly in HCI-related research techniques. Through their leadership and assistance in various projects, I was able to set new research goals and explore numerous research ideas and directions.

I am also truly grateful to all the members of our laboratory whose cooperation and support have made a significant contribution to this research. During the development and implementation of the study, their constructive feedback, positive attitude when sharing ideas, and genuine camaraderie were constructive. My heartfelt thanks go especially to those who offered assistance and encouragement whenever I needed it, creating a cooperative and motivating environment.

Finally, I wish to convey my sincerest thanks to my family, especially my parents, whose encouragement and unwavering support have provided constant strength and motivation. Their belief in my abilities helped me overcome various challenges in a foreign land, and they have witnessed every joy and hardship during this journey. Without their companionship and support, I would not have been able to achieve the milestones I have reached today.

References

[1] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative ai," *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, 2024.

[2] W. Xu, M. J. Dainoff, L. Ge, and Z. Gao, "From human-computer interaction to human-ai interaction: new challenges and opportunities for enabling human-centered ai," *arXiv preprint arXiv:2105.05424*, vol. 5, 2021.

[3] O. Ozmen Garibay, B. Winslow, S. Andolina, M. Antona, A. Bodenschatz, C. Coursaris, G. Falco, S. M. Fiore, I. Garibay, K. Grieman *et al.*, "Six human-centered artificial intelligence grand challenges," *International Journal of Human–Computer Interaction*, vol. 39, no. 3, pp. 391–437, 2023.

[4] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[5] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How much can clip benefit vision-and-language tasks?" *arXiv preprint arXiv:2107.06383*, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[9] Y. Zhang, T. Zhang, and H. Xie, "Texcontrol: Sketch-based two-stage fashion image generation using diffusion model," *arXiv preprint arXiv:2405.04675*, 2024.

[10] C. Li, T. Zhang, X. Du, Y. Zhang, and H. Xie, "Generative ai models for different steps in architectural design: A literature review," *Frontiers of Architectural Research*, 2024.

[11] Z. Huang, H. Xie, T. Fukusato, and K. Miyata, "Anifacedrawing: Anime portrait exploration during your sketching," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.

[12] X. Xie, X. Du, M. Li, X. Yang, and H. Xie, "Diffobi: Diffusion-based image generation of oracle bone inscription style characters," in *SIGGRAPH Asia 2024 Technical Communications*, ser. SA '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3681758.3698005

[13] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–23.

[14] V. Liu, H. Qiao, and L. Chilton, "Opal: Multimodal image generation for news illustration," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3526113.3545621

[15] R. Huang, H. Lin, C. Chen, K. Zhang, and W. Zeng, "Plantography: Incorporating iterative design process into generative artificial intelligence for landscape rendering," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–19.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023. [Online]. Available: https://doi.org/10.1145/3560815

[17] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

[18] C. Mitra, B. Huang, T. Darrell, and R. Herzig, "Compositional chain-of-thought prompting for large multimodal models," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14 420–14 431.

[19] Z. Shi, S. Gao, X. Chen, Y. Feng, L. Yan, H. Shi, D. Yin, Z. Chen, S. Verberne, and Z. Ren, "Chain of tools: Large language model is an automatic multi-tool learner," *arXiv preprint arXiv:2405.16533*, 2024.

[20] D. Kepel and K. Valogianni, "Autonomous prompt engineering in large language models," *arXiv preprint arXiv:2407.11000*, 2024.

[21] J. Oppenlaender, R. Linder, and J. Silvennoinen, "Prompting ai art: An investigation into the creative skill of prompt engineering," *International Journal of Human–Computer Interaction*, pp. 1–23, 2024.

[22] M. Sharma, D. Biros, C. Baham, and J. Biros, "What went wrong? identifying risk factors for popular negative consequences in ai," *AIS Transactions on Human-Computer Interaction*, vol. 16, no. 2, pp. 139–176, 2024.

[23] F. Cabitza, A. Campagner, and C. Simone, "The need to move away from agential-ai: Empirical investigations, useful concepts and open issues," *International Journal of Human-Computer Studies*, vol. 155, p. 102696, 2021.

[24] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[25] A. F. Blackwell, "Interacting with an inferred world: The challenge of machine learning for humane computer interaction," in *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*, 2015, pp. 169–180.

[26] K. Vaccaro, K. Karahalios, D. K. Mulligan, D. Kluttz, and T. Hirsch, "Contestability in algorithmic systems," in *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 2019, pp. 523–527.

[27] R. H. Harper, "The role of hci in the age of ai," *International Journal of Human–Computer Interaction*, vol. 35, no. 15, pp. 1331–1344, 2019.

[28] S. Brade, B. Wang, M. Sousa, S. Oore, and T. Grossman, "Promptify: Text-to-image generation through interactive prompt exploration with large language models," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3586183.3606725

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[30] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," *arXiv preprint arXiv:1511.02793*, 2015.

[31] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[32] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[33] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[34] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3501825

[35] A. Radford, "Improving language understanding by generative pre-training," 2018.

[36] L. Qu, S. Wu, H. Fei, L. Nie, and T.-S. Chua, "Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 643–654. [Online]. Available: https://doi.org/10.1145/3581783.3612012

[37] J. J. Y. Chung and E. Adar, "Promptpaint: Steering text-to-image generation through paint medium-like interactions," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3586183.3606777

[38] Z. Wang, Y. Huang, D. Song, L. Ma, and T. Zhang, "Promptcharm: Text-to-image generation through multi-modal prompting and refinement,"

in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642803

[39] J. Wang, Z. Liu, L. Zhao, Z. Wu, C. Ma, S. Yu, H. Dai, Q. Yang, Y. Liu, S. Zhang, E. Shi, Y. Pan, T. Zhang, D. Zhu, X. Li, X. Jiang, B. Ge, Y. Yuan, D. Shen, T. Liu, and S. Zhang, "Review of large vision models and visual prompt engineering," *Meta-Radiology*, vol. 1, no. 3, p. 100047, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950162823000474

[40] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[41] T. Wu, M. Terry, and C. J. Cai, "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–22.

[42] Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein, "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[43] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[44] T. Cao, C. Wang, B. Liu, Z. Wu, J. Zhu, and J. Huang, "Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis," *arXiv preprint arXiv:2311.06752*, 2023.

[45] H. Kim, H. Lee, S. Pang, and U. Oh, "Prompirit: Automatic prompt engineering assistance for improving ai-generated art reflecting user emotion," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2024, pp. 138–143.

[46] Y. Feng, X. Wang, K. K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen, "Promptmagician: Interactive prompt engineering for text-to-image creation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 295–305, 2024.

[47] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[51] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[52] X. Zhou, Y. Zhang, L. Cui, and D. Huang, "Evaluating commonsense in pre-trained language models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 9733–9740.

[53] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International conference on machine learning*. PMLR, 2015, pp. 2152–2161.

[54] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[55] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 267–10 276.

[56] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[57] AUTOMATIC1111, "Stable diffusion web ui," https://github.com/AUTOMATIC1111/stable-diffusion-webui.

[58] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Computer Vision–ECCV*

2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 662–679.

[59] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British journal of applied science & technology*, vol. 7, no. 4, pp. 396–403, 2015.

[60] J. Brooke, "Sus: A quick and dirty usability scale," *Usability Evaluation in Industry*, 1996.

[61] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

Appendix

Figure 10.1: Study Environment

Formative Interviews

P1(Novice users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I primarily interact with image-generation AI using text-to-image models in my daily routine, particularly leveraging DALL·E integrated within ChatGPT. Specifically, I input various themes and ideas as text prompts and enjoy generating images based on these inputs. My primary purpose for using these tools is entertainment, finding pleasure in the creative process of visualizing different concepts. Additionally, I occasionally use image-generation AI to gain design inspiration or as reference material for art projects, integrating it into leisurely activities and creative workflows.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

I typically begin with an elementary and concise description when creating prompts. For example, I might start with a basic idea like 'a sunset coast.' After reviewing the generated image, I then add more specific elements to refine the output, such as "a sunset coastline with a lighthouse standing tall." This gradual approach stems from my experience, where I found that providing complex instructions simultaneously does not guarantee that the AI will produce the desired image immediately. Instead, starting with a basic prompt and incrementally adding details based on the generated results leads to more accurate and satisfactory images. During the feedback process, I identified the strengths and areas needing improvement in the generated images. I adjusted the prompts accordingly by adding or modifying instructions to enhance the precision of the image generation.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

When the generated image does not align with my initial imagination, I first analyze the discrepancies, such as differences in color schemes, composition, or lack of detail. Based on this analysis, I make targeted modifications to the prompt. For example, if the initial prompt was "a sunset coastline with a lighthouse," and the lighthouse appears in an undesired color, I might revise the prompt to specify "a sunset coastline with a red lighthouse." Similarly, if the waves in the sea are too calm, I might adjust the prompt to "a sunset coastline with a lighthouse and turbulent waves." However, this process can be challenging because while attempting to correct the unwanted aspects, I sometimes inadvertently alter or remove desirable elements, making precise control over image generation difficult. In practice, while there are instances where the generated images significantly differ from my expectations, more often than not, the results are not entirely off-target

and retain some elements that align with my initial vision.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

Yes, I have utilized image-generation AI for creative exploration even without a clearly defined objective. In such cases, my process resembles a "gacha" or lottery system, where I input random or loosely defined text prompts and wait to see what images are produced. For example, I might enter prompts like "a futuristic cityscape" or "a mythical creature in an enchanted forest" without a specific goal. The generated images often serve as unexpected sources of inspiration, sparking new ideas I might not have conceived independently. This method allows me to discover unique and high-quality visuals that can inform future projects or artistic endeavors. The element of surprise and the potential to uncover novel concepts make this open-ended exploration enjoyable and valuable for expanding my creative horizons.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

While current image-generation AI tools offer a certain level of utility through text-based prompts, I believe they fall short of adequately supporting prompt optimization. Specifically, conveying nuanced details and subtle aspects through text alone can be challenging, often resulting in a gap between the desired image and the generated output. Historically, humanity has relied on visual mediums like images and paintings to express emotions and concepts that words alone cannot fully capture. Therefore, relying solely on textual instructions for image generation feels somewhat unnatural and may not be the most effective approach. I believe that features enabling more intuitive and direct manipulation of images without depending on language would significantly enhance the creative process.

P2(Novice users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.) Typically, I use text-to-image generation by inputting prompts to describe what I want. The AI then generates an image based on my description to match my expectations.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

Initially, I start with simple words or short sentences as prompts. If the generated result differs from my expectations, I adjust the prompt by removing un-

necessary elements, adding new details, or emphasizing certain words to better express my requirements.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

It's typical for the generated image to differ from what I envision. For example, I may not want a background in the image, but the AI generates one with a background. In such cases, I add descriptions like "no background" to the prompt to tell the AI what I don't want explicitly. However, some AI systems, like ChatGPT, struggle in this area. Even after repeated emphasis, they fail to make the necessary changes. Additionally, some image-generation AI models don't recognize those species and cannot produce accurate results when generating specific rare species.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

I usually turn to image generation AI when I have a clear idea or specific need, rather than using it casually or randomly.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

I feel that image generation AI still has shortcomings in understanding language. Sometimes, it feels like communicating with an unresponsive entity. Its ability to comprehend language is insufficient, and it often struggles with understanding specific words or expressions, making it hard to grasp my intentions accurately.

P3(Advanced users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I typically generate the required images in my daily activities by directly inputting text prompts. For example, when I need to teach AI-related content to students, I might require an illustration to demonstrate AI's generation capabilities. I use ChatGPT's image generation functionality to create PPT visuals to produce images that meet my instructional needs. For instance, I might input a prompt like "a beautiful girl in a white dress," and the AI will generate an image based on that description. This method allows me to obtain the necessary visual materials quickly and enables flexible adjustments according to specific requirements, ensuring the vividness and attractiveness of the teaching content.

Additionally, this approach saves time that would otherwise be spent searching for suitable images, enhancing teaching efficiency.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

When creating prompts, I typically start by emulating the descriptive level of sample prompts and writing a basic prompt. Subsequently, based on the generated outcomes, I continuously added more descriptions to modify elements such as the background, the actions of the main characters, and the painting style. For example, the initial prompt might only describe the basic features of the main character, such as "a beautiful girl in a white dress." Then, I gradually add details about the background environment, the character's posture, and the overall artistic style, like "a beautiful girl in a white dress waving on a sunset beach." This iterative refinement process ensures that the AI-generated images better align with my expectations and requirements. Additionally, through repeated experimentation and adjustments, I can better understand how to guide the AI to produce more accurate and high-quality images by refining the prompts.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

When the generated image does not align with my imagination, I typically add supplementary information to specific elements while introducing some negative prompt words to reinforce the direction of generation. For example, I might input "a red bird instead of a green one" to prompt the AI to change the bird's color to red and ensure that no green birds appear in the image. Generally, the generated images are mainly consistent with the prompt specifications, but some details, such as the background color and the main subject's position., may be inconsistent. I must add prompt words to adjust these finer details in such cases. Additionally, some aspects within the image may not meet my expectations; in such instances, I identify these discrepancies and specify them in the prompt to guide the AI in making the necessary modifications. While such situations do not occur frequently, appropriate adjustments and feedback are essential to ensure that the generated images fully meet my expectations.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

I conduct such attempts infrequently because the effectiveness of AI-generated content under specified prompts still needs improvement. However, I once tried

asking the AI to create an image of how it perceives me based on the questions I had previously posed. The purpose of this attempt was to observe what kind of image my daily inquiries would shape. Since I frequently asked AI code-related questions during that period, the AI generated an image of an engineer typing code in front of a computer, which I found entirely accurate. Although limited, this attempt showcased the AI's potential in specific tasks. Additionally, without a clear generation objective, I occasionally input random or vague prompts to see what images the AI would produce, hoping to discover new creative inspirations.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

In terms of support, I believe enhanced control over layers could be added. For instance, the ability to generate specific content in designated areas of the image while ensuring that the entire image aligns with the corresponding prompt content would be beneficial. Additionally, incorporating functionalities for local editing of already generated images, such as adding perspective or camera distance conditions, could create more suitable and precise images. By enhancing these features, users can conduct creative explorations more effectively and make finer adjustments and optimizations to the generated images, thereby improving the overall image generation quality and better meeting specific requirements.

P4(Advanced users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I typically interact with image-generation AI using text prompts and sketches. Text prompts are my primary method for generating images and guiding video outputs. Sketches, however, help provide additional visual references, allowing the AI to generate results that better match my specific requirements.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

My process usually starts by describing the core content I want to generate, focusing on key nouns to establish the foundation. Once the essential content is precise, I add descriptive words to refine the style or tone. For instance, I might adjust prompts with specific adjectives or stylistic terms to ensure the output aligns with my vision. Based on the feedback from the generated results, I refined the prompt step by step, adding missing details or modifying elements that didn't meet my expectations.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

When the results don't meet my expectations, I adjust the seed value and try several variations. If trying around ten seeds doesn't produce satisfactory results, I modify adjectives tied to key nouns or restructure the main components of the prompt, such as replacing or rephrasing verbs. It's common for the generated images to deviate from what I imagined, especially when the prompts lack sufficient detail or clarity.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

No, I usually use image-generation AI with specific objectives in mind. My interactions are goal-driven, focusing on generating images aligned with a clearly defined purpose rather than open-ended creative exploration.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

Current tools are not sufficient for prompt optimization. Integrating features that expand simple user inputs into detailed and structured prompts would be helpful. For instance, using a text-to-text model (like ChatGPT) to supplement user-provided descriptions with additional details could make the process more efficient. Users could then review and adjust these expanded prompts. Additionally, having examples or templates categorized by use cases could simplify the creation process and make it easier for users to craft prompts that align with their desired outcomes.

P5(Advanced users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I primarily use generative tools in two ways: by directly operating at the algorithmic level and through visual interfaces like ComfyUI. For input data, I utilize various types of auxiliary image information such as depth, sketch, image, edge, and normal maps to enhance the generation process. My main focus is fashion-related content generation, including realistic and anime-style images. Additionally, I'm exploring tasks related to video generation. In video generation, I typically break the task into generating a sequence of consistent frames, such as producing 120 images with coherent motion, to achieve smooth video output.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

I distill the final target into a concise sentence supplemented by key descriptors. For instance, if I aim to generate an image of an evening dress, I might start with a simple phrase like "A photo of an evening dress." Alongside this, I include a set of positive and negative descriptors. Positive descriptors might include "high quality" and "best quality." In contrast, negative descriptors could include terms such as "long body," "low-res," "bad anatomy," "fewer digits," "cropped," "worst quality," and "low quality." These descriptors are usually fixed based on my prior experiments and have been proven effective in producing the desired images with certain seeds. Once these foundational elements are established, I optimize the network structure. However, when it is evident that modifying prompts significantly impacts the results, I also add or remove specific phrases accordingly. Overall, my adjustments to prompts are primarily focused on extending or refining the core sentence.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

When results deviate from expectations, I typically check if using synonyms might have caused CLIP to misinterpret the prompt. For example, when generating clothing with a fur-like texture, using "fur" alone might create ambiguity, as "fur dress" could be misinterpreted as animal fur. In such cases, I prefer more specific terms like "fur cloth" to avoid confusion. Predefined prompts usually work without significant issues for single-object generation. However, even with precise text prompts and sketch-based assistance, inconsistencies in the output can still occur for multi-object generation.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

I do not use generative tools without a clear goal. Instead, I primarily employ them to explore generation quality and test controllability, ensuring the models perform as intended.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

I rely on prompts and predefined outputs to validate my generative algorithms. Through this process, I've observed that better prompts can lead to significantly

improved results. Thus, having high-quality preset prompts and prompt suggestion functionalities is highly beneficial. Furthermore, in experiments involving sketch and image-based assistance, analyzing these inputs to suggest potentially applicable prompts could be a practical and effective improvement.

P6(Intermediate users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I mainly engage with these tools out of curiosity and experimentation rather than aiming for a specific creative outcome. Sometimes, I try using DALL to see what it can generate; at other times, I explore various web-based image generation services (whose names I've unfortunately forgotten). This carefree approach allows me to understand each tool's strengths and results.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

When creating prompts, I adhere to a "keep it simple" principle—usually just one concise sentence, without piling on too many attributes or details. If the initial output doesn't quite match what I had in mind, I tweak that same prompt slightly, experimenting with different keywords or phrasing to see how the model's response changes.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

Almost every generated image differs from my initial mental picture. When encountering this mismatch, I prefer introducing different keywords or concepts rather than making the prompt longer or more complex. Unnecessary layers of description seldom lead to a more satisfying result.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

So far, I haven't experimented with these tools in a state of having absolutely no direction at all. Typically, I like to have at least a rough idea before I start.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

In an ideal scenario, I'd like to adjust the intensity or prominence of specific image attributes using something more intuitive than pure text—perhaps a set of

multidimensional sliders. If the image could update in real-time as I manipulate those sliders, I'd gain a far more direct and immediate sense of how each adjustment influences the final output.

P7(Intermediate users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I begin by describing the image I have in mind to ChatGPT, prompting it to generate a set of instructions. ChatGPT's output is sometimes overly detailed and surpasses Stable Diffusion's token limit of 77. I manually refine and shorten the prompt to fit SD's constraints in such cases. If text-based generation via SD fails to achieve the desired results, I turn to more advanced tools like ControlNet, allowing me to input a rough sketch alongside the prompt. This visual reference helps guide the composition and style more effectively. Suppose the outcome remains unsatisfactory even with a sketch. In that case, I'll continually modify the sketch, effectively engaging in an iterative, real-time feedback loop—adjusting my drawing and regenerating the image until it aligns closely with what I envision.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

I start with simple English prompts: "Old man, long hair, with glasses." Since the chosen model often performs better with English descriptions, I'll rely on translation tools to produce a basic English draft of the prompt if my language skills are insufficient to convey a complex idea. If I'm still unsatisfied with the outcome, I'll input that draft into ChatGPT and request a more polished, finely tuned prompt that better captures the nuances of my intentions.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

To refine the generated images, I add more descriptive keywords, specifying elements like art style (e.g., "anime style") or particular details (e.g., "wrinkles on the forehead").

Regarding highly specialized or professional imagery, the model may inherently lack the capability to render such content accurately. Even after multiple prompt adjustments, the improvements might remain marginal, and it is challenging to reach a satisfying result. On the other hand, for more commonplace imagery—like a generic human figure—where I have few strict criteria, it's relatively easy to feel content with the generated output. In these less demanding scenarios, I don't have

a clear standard for what would be considered "unsatisfactory," so I'm usually okay with what the model provides.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

I often use DALL·E for open-ended exploration when I have no apparent objective. One of DALL· E's strengths is that it can provide suggestions or pose questions after generating an image. This interactive process helps me gradually discover and clarify what I want as I respond to DALL· E's guidance and refine my goals.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

From my perspective, ChatGPT's capabilities in generating prompts are already reasonably sufficient. They support me in starting and effectively refining my creative process.

P8(Advanced users)

1. How do you typically interact with image-generation AI? (Methods, purposes, etc.)

I primarily interact with image-generating AI in three ways. First, I provide textual prompts to guide the AI in producing the images I want. Second, I start with an existing image and specify a region I'd like to modify using additional prompts, enabling me to refine the image until it meets my requirements. Third, I sometimes provide a simple sketch alongside prompts, offering the AI a visual reference. This helps ensure the generated image aligns more closely with my initial vision.

2. What steps do you normally take when creating prompts? For example, do you start with simple descriptions and gradually add detail? Also, how do you incorporate feedback from the generated results into improving your prompts?

My general approach to crafting prompts is incremental. I begin with core keywords or phrases that capture the main idea and then gradually add more detailed, specific terms. Loading all details into a single prompt from the start is often time-consuming and inefficient, as the initial outcome may not match my desired vision. Instead, I prefer to generate multiple samples using more minor prompts. I select the one closest to my intended direction from these samples and refine it with additional, more detailed prompts. This step-by-step method helps me maintain better control over the creative process.

3. How do you modify your prompt When the generated image differs from what you imagined? Also, do you frequently encounter situations where the image produced is quite different from your initial conception?

If the generated image doesn't match my expectations, I identify which elements are off-target or missing. If the result includes irrelevant or unwanted details, I introduce negative prompts to discourage their appearance in subsequent iterations. For elements that fail to appear as intended, I add or strengthen specific prompts and possibly adjust their weighting to ensure they appear more clearly. Adjusting the prompt this way is standard, as misalignments between the AI output and my initial vision occur frequently.

4. Have you ever used image-generation AI for idea generation or creative exploration without a clear objective? If so, could you describe your process?

Yes, I've experimented with image-generating AI without a clearly defined goal, using it as a tool for creative exploration and inspiration. For example, I might start by specifying a general visual style—such as a realistic or anime-inspired look—and essential characteristics like the subject's gender, whether a full-body portrait or a close-up. From there, I observe the initial results and refine them, step by step, adding detail: facial features, hairstyle, expression, accessories, clothing if it's full-body, and even poses. Throughout this iterative process, I carefully assess the AI's output, using each round of generation to guide further prompt refinement.

5. Do you believe current tools sufficiently support prompt optimization when using image-generation AI? If not, what features would make creative exploration more effective?

Current tools for prompt optimization still leave room for improvement. For instance, right now, adjusting the weight of a prompt often relies on specific symbols or syntax. A more intuitive method—like using percentage sliders—would make fine-tuning more transparent and user-friendly. Additionally, organizing prompts into categories and offering example keywords for each would help streamline the selection process and reduce the user's cognitive loaIfIfystemtem could automatically suggest a range of more granular, nuanced prompts upon entering a broad keyword, it would make it much easier for users to pinpoint terms that precisely match their creative intentions.

NASA-TLX Questionnaire

Instructions: Please rate each dimension on a scale of 1 to 7 by selecting the appropriate box.

Dimension	1	2	3	4	5	6	7
Mental Demand: How mentally demanding was the task?							
Physical Demand: How physically demanding was the task?							
Temporal Demand: How hurried or rushed was the pace of the task?							
Performance: How successful were you in accomplishing what you were asked to do?							
Effort: How hard did you work to accomplish your level of performance?							
Frustration Level: How insecure, discouraged, irritated, stressed, and annoyed were you?							

Table 10.1: NASA-TLX Questionnaire

System Usability Scale (SUS)

1. I want to use this system frequently.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

2. I found the system unnecessarily complex.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

3. I thought the system was easy to use.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

4. I think that I would need the support of a technical person to use this system.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

5. I found the various functions in this system were well integrated.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

6. I thought there was too much inconsistency in this system.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

7. I would imagine that most people would learn to use this system very quickly.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

8. I found the system very cumbersome to use.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

9. I felt very confident using the system.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

10. I needed to learn a lot of things before I could get going with this system.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Table 10.2: System Usability Scale (SUS)

Subject Rating

1. I feel that the final generated image is visually appealing.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
2. The content presented in the image (details, elements, ambiance, etc.) is rich and aesthetically pleasing.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
3. The key elements from previous prompts/images have been effectively integrated into the final image.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
4. The final generated image adequately reflects my expectations for the key elements or style.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
5. The overall style of the generated image is consistent with the theme/concept I initially envisioned.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
6. Even after multiple iterations, the image style remains fairly coherent and consistent.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
7. I can see an improvement or enhancement in the final image compared to previous iterations.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
8. Through multiple iterations, the final image better aligns with my ideal outcome.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
9. The system provides sufficient flexibility for me to explore different directions or variations during the creative process.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
10. I can easily incorporate new ideas or elements while retaining previous advantages, allowing me to explore more possibilities.
☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Table 10.3: Subject Rating