

PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation

Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen

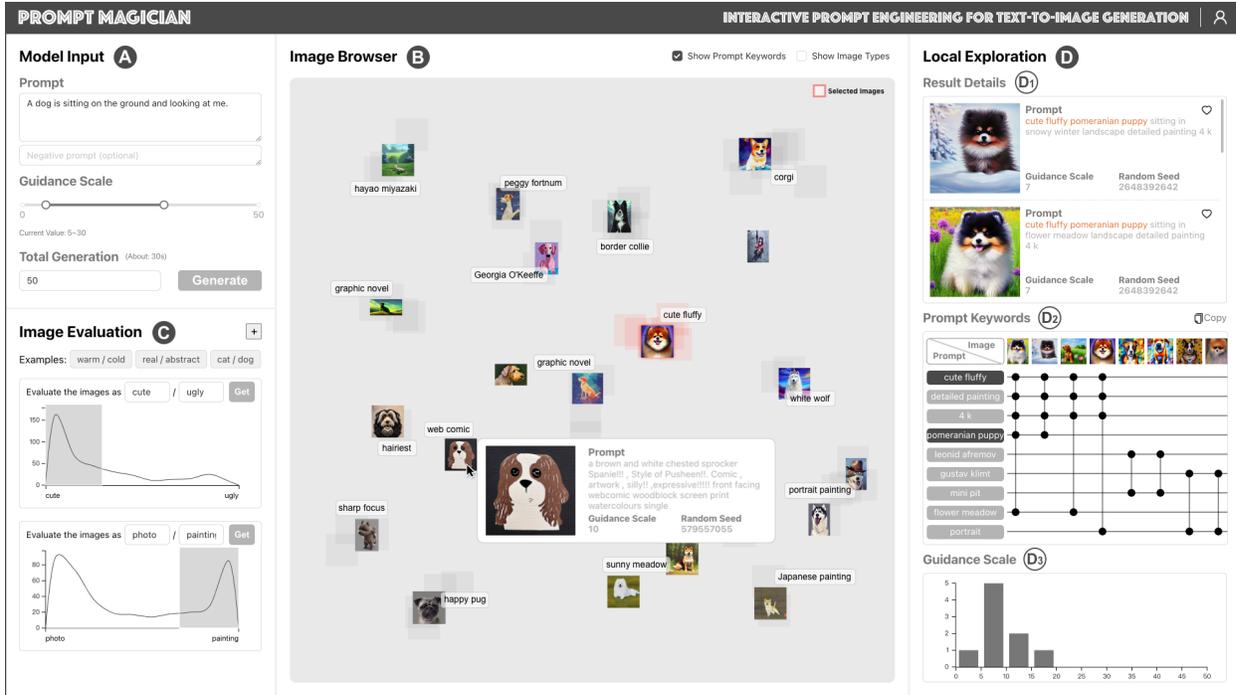


Fig. 1: The user interface of *PromptMagician* consists of four views. The *Model Input View* (A) configures the prompts and hyper-parameters for image creation. The *Image Browser View* (B) visualizes the generated and retrieved images and the recommended prompt keywords. The *Image Evaluation View* (C) helps evaluate and filter images based on multiple criteria. The *Local Exploration View* (D) helps users explore and validate the prompt keywords and guidance scales for images of interest.

Abstract—Generative text-to-image models have gained great popularity among the public for their powerful capability to generate high-quality images based on natural language prompts. However, developing effective prompts for desired images can be challenging due to the complexity and ambiguity of natural language. This research proposes *PromptMagician*, a visual analysis system that helps users explore the image results and refine the input prompts. The backbone of our system is a prompt recommendation model that takes user prompts as input, retrieves similar prompt-image pairs from DiffusionDB, and identifies special (important and relevant) prompt keywords. To facilitate interactive prompt refinement, *PromptMagician* introduces a multi-level visualization for the cross-modal embedding of the retrieved images and recommended keywords, and supports users in specifying multiple criteria for personalized exploration. Two usage scenarios, a user study, and expert interviews demonstrate the effectiveness and usability of our system, suggesting it facilitates prompt engineering and improves the creativity support of the generative text-to-image model.

Index Terms—Prompt engineering, text-to-image generation, image visualization

1 INTRODUCTION

Generative text-to-image framework has become a popular and effective interactive paradigm [44] with widespread adoption in academia [33, 42, 43, 78] and the public [23, 58]. The endless space of natural language text allows for the free expression of artistic ideas and significantly lowers the barrier to image creation. With the rapid development of natural language processing (NLP) and computer vision (CV) technologies, state-of-the-art generative models, such as Stable Diffusion [43] and DALL-E 2 [42], have been able to generate relevant and high-quality images based on text prompts and have demonstrated great potential in downstream tasks, including hyper-realistic video generation [17] and radiology image synthesis [5].

Building upon the success of these generative models, researchers and developers have explored a human-model interaction technique

arXiv:2307.09036v2 [cs.AI] 15 Aug 2023

- Y. Feng, S. Wang, Y. Lu, B. Wang, and W. Chen are with the State Key Lab of CAD&CG, Zhejiang University. W. Chen is also with the Laboratory of Art and Archaeology Image (Zhejiang University), Ministry of Education. W. Chen is the corresponding author. Email: {fycj, sjiawang, luyuhong, wangbaicheng, chenwis}@zju.edu.cn.
- X. Wang and KK Wong are with the Hong Kong University of Science and Technology. Email: {xwangec, kkwongar}@connect.ust.hk.
- M. Zhu is with Zhejiang University. Email: minfeng_zhu@zju.edu.cn.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxx

called “prompting” [14, 45, 59]. During the creation process, users craft natural language prompts that describe the expected image characteristics (e.g., subjects and styles), adjust model hyper-parameters (e.g., guidance scales) and try out more seeds to obtain the desired output. However, the complexity and ambiguity of natural language can make it challenging for users, especially novice users, to develop effective prompts that trigger the model to generate the desired output [54]. Additionally, prompts can result in distinct images based on different model hyper-parameters. It is difficult to evaluate the quality of the prompts with the limited trial of hyper-parameter values. When receiving undesirable image results, users may become confused about whether and how to adjust the prompts or model hyper-parameters. Previous research has proposed automatic prompting techniques [57] for text-to-image creation. However, the image creation process greatly depends on human subjective judgment, which requires humans in the loop to refine the generation. Some research [36, 58] suggests using “magical spells” (e.g., keywords) to formulate prompts based on large human-annotated corpora. However, these guidelines could be too general to satisfy personalized image creation needs.

To address these challenges, we present *PromptMagician*¹, a novel visual analysis system for interactive prompt engineering. It aims to help users efficiently explore and evaluate the model-generated images and refine the input prompts and hyper-parameters for the desired image results. Given a text prompt, the system automatically generates a collection of image results with a range of hyper-parameter values and retrieves related prompt-image pairs from DiffusionDB, a large prompt-image corpus [58]. Then, the system presents a visual summary of both the generated and retrieved images to guide the exploration of images with diverse styles. Additionally, users can specify image evaluation criteria using descriptive words (e.g., “good” for image quality and “beautiful” for abstract perception) to filter out irrelevant images and focus on the image subset of interests for efficient exploration.

To assist users in prompt improvement, we propose a prompt keyword recommendation model based on prompt engineering design guidelines [23], which prioritizes prompt keywords over sentence structures. The model encodes the retrieved images from DiffusionDB with the CLIP model [41], organizes them into hierarchical clusters, and identifies special (important and relevant) prompt keywords from the corresponding prompts of the image clusters. The importance of the keywords is measured using cluster-level TF-IDF values [49]. Finally, the model matches the keywords with their most related clusters and recommends them to users for prompt improvement. The recommended prompt keywords are visualized alongside the matched image clusters to facilitate user exploration. The users can select image subsets to explore their prompt keywords and guidance scale. The system also visualizes the interrelationship between the prompt keywords and images to help users understand and compare the effects of different prompt keywords for image creation. We evaluate our system and prompt recommendation model through two usage scenarios, a user study, and expert interviews, and the results show that our system can help users discover effective prompt keywords and inspire image creation.

In summary, our major contributions include:

- A visual system to help users explore and evaluate the model-generated results and conduct interactive prompt engineering.
- A prompt recommendation model that identifies important and relevant prompt keywords to help prompt improvements.
- Two usage scenarios, a user study, and expert interviews that demonstrate the effectiveness and usability of our system.

2 RELATED WORK

2.1 Prompt Engineering

With the rapid development of large language models [3, 37] and text-to-image models [33, 42, 43], prompt engineering [14, 45, 59] has become a promising paradigm for interacting with models [20]. With this paradigm, users can focus on designing and refining the prompt input

¹The code is available at <https://github.com/YingchaojieFeng/PromptMagician>

to improve the performance of the pre-trained model in specific application scenarios, directly utilizing the knowledge and capability of the pre-trained model without the additional training process. Nowadays, it has gained widespread attention and shown great potential in various tasks, such as natural language understanding [19], image generation [23], and logical reasoning [61].

Previous studies have focused on automatic approaches for prompt formulation and refinement. AutoPrompt [46] applied gradient-guided search in the collection of trigger tokens to automatically create prompts for masked language models. Gao *et al.* [14] employed the generative T5 model to generate the prompt templates and pruned brute-force search for label word selection. To facilitate human-AI collaboration in prompt engineering [62], interactive and visual systems were proposed. PromptIDE [51] provides interactive visualizations to help users evaluate the performance of prompts on a small dataset and iteratively refine prompts. For complex tasks that require multi-step operations, PromptChainer [61] allows users to interactively construct chains of prompts for the corresponding targeted sub-tasks, increasing the transparency and controllability of large language models.

Most of the aforementioned studies are designed for text-to-text generative models whose output can be transformed into label results and used for the quantitative evaluation of prompt performance on a given dataset. Our work focuses on text-to-image generative models, which have different outputs and evaluations [57]. To provide guidelines for prompting research, Liu *et al.* [23] conducted experiments to explore a set of open questions in prompt engineering for text-to-image models. The results emphasized the importance of the prompt keywords (*i.e.*, subject and style) over the phrasing structures. Based on an ethnographic study with community practitioners, Oppenlaender [35] summarized a taxonomy of prompt modifiers, including subject terms, modifiers, and magic terms, to guide and inspire prompt formulation. Nevertheless, these guidelines could be too general to satisfy personalized image creation needs. A recent work, RePrompt [57], introduces explainable AI techniques (e.g., SHAP value [28]) to reveal the importance of text features, including the numbers and concreteness of each POS (part of speech) type, and their optimal value ranges. Opal [24] utilized GPT-3 [3] to generate text prompts for new illustrations. Our work differs from prior work by combining database retrieval and ad-hoc generation, enabling users to explore the vast artistic search space to identify effective prompt keywords for personalized creation and iteratively refine the prompts.

2.2 Visual Exploration of Image Collections

A large number of daily-created images provide rich information for various applications, such as AI model development [6] and content retrieval [64]. Prior studies have proposed many techniques for image exploration at scale, such as tree-based visualizations [2], enhanced scatter plots [55, 60, 63, 77], and node-link graphs [21, 68, 69].

One major challenge is the summary and exploration of the complex semantics associated with images [16, 73]. Semantic Image Browser [65] annotates semantic content in images and uses a Multi-Dimensional-Scaling-based image layout that aggregates semantically similar images together. Similarly, Xie *et al.* [64] produced semantic image descriptions using image captioning techniques. Then, they utilized a co-embedding model to project images and their semantic descriptions into the same 2D space and employed a galaxy metaphor to provide a semantic overview of image collections. Chen *et al.* [6] proposed a node-link-based visualization powered by a co-clustering algorithm to reveal object labels and images in object detection tasks. The interactive visualizations help users explore and validate labels of the detected image objects. Compared to prior work that mostly concentrates on image content exploration, we further consider other factors, such as image styles and model hyper-parameters, to help users formulate and refine their prompts to create visually appealing images using generative models. For example, we use a pre-trained vision-language model, CLIP, to encode images, which considers both image content and visual styles. Based on model encodings, our system allows users to evaluate and filter images using natural language descriptions about image properties, such as “cartoon” and “beautiful.”

2.3 Text-to-Image Generation

Text-to-image generation refers to translating natural language descriptions (e.g., words and sentences) into realistic images. Recent breakthroughs in computer vision (CV) and natural language processing (NLP) techniques have greatly improved text-to-image generation quality. Modern text-to-image models typically utilize an encoder-decoder architecture, where encoders learn the contextual representations of input text, and decoders use the learned information to generate corresponding images. Particularly, text encoders are usually pre-trained language models (e.g., GPT [3] and BERT [10]), and to-image decoders generally use GAN-based and Diffusion-based models.

GAN-based models [70, 78] contain a generator and a discriminator, where the generator accepts text encodings and generates output while the discriminator tries to differentiate the output from real image examples. Diffusion-based models [33, 42, 43] learn to remove noise from random images and generate final images that match the text information. For instance, Stable Diffusion [43] involves a latent diffusion process where the model learns to remove noise from the random noised images in the embedding space with the guidance of text input. The denoising process leads to high-quality images with state-of-the-art performance.

However, text-to-image generation quality greatly depends on natural language prompts and human subjective judgment. It requires humans in the loop to refine the generation [76]. Although there are some open-sourced demos, such as Stable Diffusion², Midjourney³, and DALL-E 2⁴, for the public to create their own artwork with natural language input, users need to try different phrasings to derive the desired output, which can be time-consuming. In this paper, we propose an interactive visual analytics system that can summarize and recommend prompt keywords to help formulate and refine users' prompts based on an external large text-to-image prompt dataset, DiffusionDB [58].

3 OVERVIEW

3.1 Background

Stable Diffusion. Based on the observations of particle diffusion in physical systems and modeling of the inverse process [47], denoising diffusion probabilistic models achieved significant improvement in generating high-resolution images [18, 48]. For diffusion models, a forward process is defined by a series of steps for adding noise to the image, and the corresponding backward process (i.e., the denoising process) is modeled by deep neural networks. The denoising process can be guided by extra conditions, such as text inputs (i.e., text prompts). The importance of the text prompt guidance is controlled by the hyper-parameter *guidance scale*. A larger guidance scale brings better alignments between the generated images and the prompts with the sacrifice of image diversity. Stable Diffusion [43], one of the state-of-the-art diffusion models, achieves high performance while consuming fewer computational resources. By compressing the images from pixel space into latent space, Stable Diffusion preserves the semantic information while removing the image details, resulting in a simplified representation space and a faster generation process. Our study utilizes Stable Diffusion for image generation, and it can be replaced by other text-to-image generative models for specific applications.

DiffusionDB. The popularity of Stable Diffusion has led to a surge in individuals sharing their image creations and input prompts on public social platforms. This trend has sparked new studies aimed at collecting and analyzing publicly shared results for future research opportunities. DiffusionDB [58] is the first large-scale dataset that comprises 14 million input-output data pairs (i.e., text prompts and hyper-parameters input by users and their corresponding model-generated images). DiffusionDB anonymizes image creators to protect user privacy and excludes harmful or NSFW (not safe for work) images. Since some users may use the same text prompt for several attempts with different

hyper-parameters (e.g., random seeds and guidance scales), the 14 million data items contain 1.8 million unique text prompts. To facilitate prompt feature analysis, DiffusionDB also offers a subset version called DiffusionDB-2M, which includes 1.5 million unique text prompts and their corresponding 2 million generated images.

3.2 Design Requirements

The target users of our study are ordinary users who are interested in image creation but lack the expertise to use professional tools. Often, these users struggle to produce high-quality images that meet their expectations. A potential solution to this problem is the utilization of a text-to-image generative model. The primary objective of our study is to design a system that facilitates collaboration between the users and the generative model. We recruited 9 ordinary users (P1-P9) who are interested in text-to-image creation from local universities. P1-P4 are familiar with interactive tools or online demos for image creation, and others are aware of such tools but are not very familiar with them.

To identify the design requirements of the system, we interviewed the participants during the early stage of the study. In the interview, participants were asked to create images using publicly available systems, including the online demo of Stable Diffusion and Lexica⁵, an online website system that supports similar image-prompt search. We encouraged participants to engage in open-ended exploration without the constraints of content or style. Then, we conducted interviews with participants to collect their feedback and comments regarding the usage of text-to-image models. Specifically, we focused on how the participants get the expected results and how to facilitate this process. In the following four months, we conducted regular meetings with them to update the design requirements and gain feedback to guide the design and development of our system prototype. Finally, the design requirements of our system were summarized as follows.

R1. Generate a collection of image results for the user prompt.

In some online demos of text-to-image tools, the users can only get a few results per prompt submission and have to manually try different model inputs (i.e., text prompt and model hyper-parameter) each time to get more image results. It may be time-consuming for the users to get the desired results. The system should help users efficiently get a collection of image results for exploration.

R1.1. Generate multiple image results with varying hyper-parameter values. For the same user prompt, the Stable Diffusion model can generate different image results using different hyper-parameters, including the guidance scale and random seed. The users often encounter confusion when receiving undesirable image results, as it is challenging to assess whether the prompt itself is inadequate or requires better model hyper-parameters. The system should allow users to specify multiple hyper-parameter values to generate multiple images at once, which helps users efficiently evaluate the quality of prompts.

R1.2. Provide similar image work to inspire prompt refinement. The target users of our system do not have to be familiar with the model's architecture and prompting strategies. By exploring and comparing the image works and their prompts, users can gain insight into what kinds of results the model is capable of generating and how to phrase or refine their prompt to obtain such results. The system should provide previous image works that are similar to user prompts and help users gain inspiration for prompt refinement.

R2. Provide a visual summary for image collection. The purpose of image exploration is to find image results of interest to inspire prompt refinement. However, exploring a large image collection is a time-consuming process. The system should provide a visual summary of the image collection so that users can easily overview the image characteristics and navigate to the image subset for detailed exploration.

R3. Support efficient image evaluation from different aspects. The images in the collection may have diverse subjects or styles. Users desire to specify evaluation criteria for images from different aspects (e.g., image subject, color style, and visual perception) and automatically evaluate the image results so that users can gain an overview of

²<https://huggingface.co/spaces/stabilityai/stable-diffusion>

³<https://www.midjourney.com/>

⁴<https://openai.com/product/dall-e-2>

⁵<https://lexica.art>

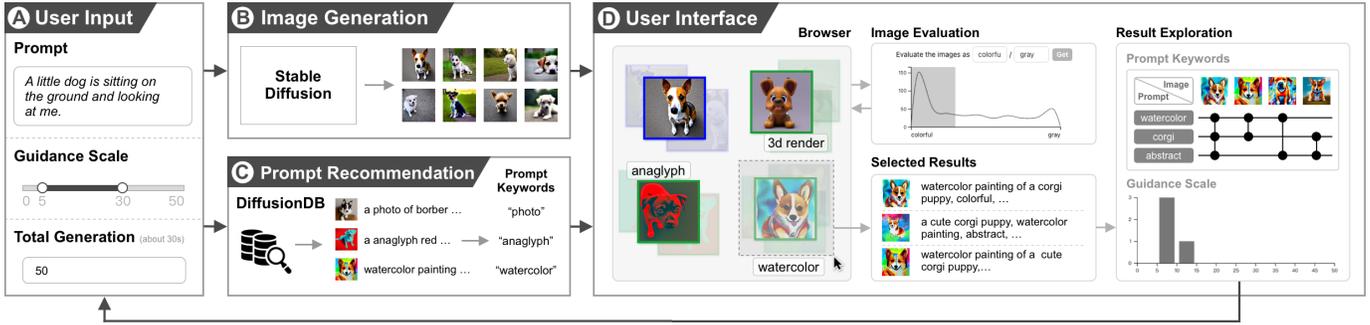


Fig. 2: The *PromptMagician* framework consists of four major components. It enables users to (A) specify model input for text-to-image creation. *PromptMagician* (B) generates a set of images using Stable Diffusion and (C) identifies related prompt keywords for recommendations. (D) Both the image results and prompt keywords are visualized in the user interface to support interactive exploration for prompt engineering.

the image distribution in terms of these aspects and focus on the image subset of interests for efficient exploration.

R4. Support iterative refinement of prompt and model hyper-parameters. Given the flexibility of natural language and the subjective nature of image creation, users usually need to iteratively refine the model input, including prompt and model hyper-parameters (*i.e.*, guidance scales and random seeds), to get the desired results. However, it is challenging to identify effective prompt keywords. The system should recommend prompt keywords for the related images and support the joint exploration of prompt keywords and their corresponding images for validation. Moreover, the system should help users explore the model hyper-parameters of the images of interest.

3.3 System Overview

The workflow of our system is summarized in Figure 2. Our system supports user input of prompts and model hyper-parameters (R1), including the range of guidance scale and the number of generations (for different random seeds). Then the system generates a collection of images using prompts and hyper-parameters (R1.1). To help users improve the prompts, the system introduces a prompt recommendation model that retrieves similar creation results from DiffusionDB (R1.2) and identifies the related prompt keywords from the corresponding prompts. Both the model-generated and retrieved images and recommended prompt keywords are co-embedded into 2D space according to their semantics and presented in multi-level visualization to facilitate exploration (R2). Based on that, the system enables users to specify the aesthetic evaluation criteria (*e.g.*, beauty) to efficiently evaluate and select the images (R3). The users can select image subsets of interest for further exploration of their prompt keywords and guidance scales, which can be used to refine the user input (R4).

4 PROMPT RECOMMENDATION

The prompt recommendation model mines special and relevant prompt keywords from similar image creations. As shown in Figure 3, the model pipeline consists of five steps: (A) retrieving image results similar to user input prompts from the DiffusionDB dataset; (B) embedding images according to their semantic features; (C) conducting hierarchical clustering of images; (D) identifying important and special prompt keywords from image clusters; and (E) matching each prompt keyword to its most related image cluster.

4.1 Image Retrieval

To retrieve similar images that match the user prompts, both images and their original prompts in the DiffusionDB dataset [58] can be used as a search space. However, due to the nature of text-to-image generative models [23], many cases in the DiffusionDB contain significantly different images generated by the same or very similar prompts [58]. Using the image features as the search space can better distinguish their differences and return more similar results. To align user prompts with the image space, we utilize the CLIP model [41], a state-of-the-art model in contrastive representation learning. Pre-trained on a vast

dataset with 400 million text-image pairs, the CLIP model has aligned the feature vectors of the text and image in each pair. We use cosine distance to measure the feature similarity of user prompts and images.

4.2 Image Embedding

For the retrieved results, we utilize both the images and their prompts for embedding to better capture the semantic features of the images, such as the subjects and style. Similar to the encoding schema for image retrieval, we employ the CLIP model to encode both the images and prompts into 512-dimensional vectors. Then, the text and image features are concatenated together as $512+512=1024$ dimensional vectors, which serve as the final representation of the images. Through this process, we can uncover the semantic similarity of images based on their distance in the 1024-dimensional space. Compared to individual text or image features, the concatenated features can better aggregate similar images with similar prompts. To enable user exploration of images according to their semantic similarity, we employ the t-SNE algorithm [52] for feature dimension reduction. We set the cosine distance as the *metric* parameter of the t-SNE algorithm, aligning it with the training objective of the CLIP model.

4.3 Hierarchical Clustering

Based on the representation features of images that reveal their semantic similarity, we organize the images with a clustering algorithm to aggregate common characteristics (*e.g.*, image colors and styles). Since there are no discrete labels available for the retrieved images, it is hard to determine the number of clusters for images, which is an important parameter that needs to be pre-specified for some clustering algorithms, like k-means [31]. Therefore, we use hierarchical clustering, which models the cluster structure as a tree. As shown in Figure 3C, each image (leaf nodes in the tree) is initialized as its own cluster and progressively merged with the neighboring nodes into a larger cluster (non-leaf nodes in the tree). The clustering process is bottom-up and ends when all the image nodes are merged into one tree.

Each non-leaf node in the tree denotes an image cluster in the embedding space, but not all clusters are suitable for prompt keyword mining, as we aim to identify the prompt keywords that are special and strongly associated with the image clusters. Overly large clusters contain many generic keywords (*e.g.*, stop words) and obscure specific ones (*e.g.*, magical words [58]). We constrain the volume of clusters based on the number and position range of child nodes. We limit the range of the number of child nodes from 3 to 20 according to the total number of retrieved images. For the positional constraints in the 2D space, we discard clusters merged from relatively distant sub-clusters, which are not appropriate for prompt keyword mining.

4.4 Prompt Keyword Mining

In line with prior studies [23, 36] that emphasize the importance of prompt keywords (*i.e.*, prompt modifiers and phrases) over connecting words (*i.e.*, sentence structure), we focus on recommending prompt keywords to help users refine the original prompt input. For each cluster,

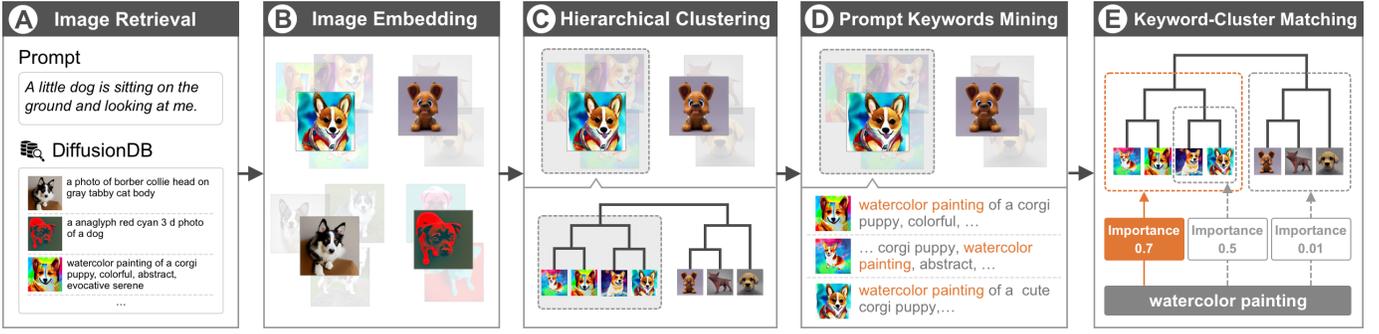


Fig. 3: The pipeline of the prompt recommendation model involves five steps: (A) retrieving similar images from the DiffusionDB dataset; (B) embedding them according to their semantics; (C) arranging them into hierarchical clusters; (D) mining special and related keywords from the prompts in the clusters; and (E) matching each keyword with its most related cluster.

we aim to identify special keywords from the prompts of the image clusters. By “special,” we mean that these keywords are significantly more crucial for the current cluster than for others. To measure the importance of each keyword for the current cluster, we compute the TF-IDF values of keywords at the cluster level:

$$\text{tfidf}_{i,x} = \text{tf}_{i,x} \times \text{idf}_i \quad (1)$$

where the $\text{tf}_{i,c}$ is the term frequency of the given keyword t_i in the current cluster c_x . It measures the importance of the keyword for the current cluster and is calculated as follows:

$$\text{tf}_{i,x} = \frac{n_{i,x}}{\sum_k n_{k,x}} \quad (2)$$

the $n_{i,x}$ is the number of keyword t_i in the current cluster c_x which consists of multiple prompt documents d_j :

$$n_{i,x} = \sum_{d_j \in c_x} n_{i,j} \quad (3)$$

The idf_i is the inverse document frequency of this keyword which measures the inverse importance of the keyword for the whole prompt set of the retrieved images.

$$\text{idf}_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (4)$$

where the $|D|$ is the total number of prompts for the retrieved image collection and $|\{j : t_i \in d_j\}|$ is the number of prompts containing the given keyword t_i . Consequently, generic keywords tend to receive higher tf values but lower idf values, which renders them less likely to be identified as the most special keywords for the cluster.

Moreover, since the special prompt keywords can comprise multiple words [58], such as “*unreal engine*” and “*trending on Artstation*,” we incorporate n -grams to detect special multi-word phrases in the prompts. After computing the importance value of the keywords in the cluster, we eliminate the stop words (e.g., “*the*” and “*and*”) using the NLTK toolkit [26]. Please note that the stop words in the n -gram (e.g., “*on*” in “*trending on Artstation*”) are excepted since they are connecting words for the other words. Finally, we select the top prompt keywords for each cluster according to their importance values.

4.5 Prompt-Cluster Matching

Following prompt keyword mining, the keywords may occur in multiple clusters, each with varying levels of importance. Matching the prompt keywords to their most related cluster node can better illustrate the effect of the keywords [74] (i.e., what can be generated by the prompt keywords). This mapping is especially beneficial when visualizing the images and text jointly to aid user exploration and comprehension [64]. To accomplish this, we normalize the importance values of keywords

within the same cluster and select the cluster with the highest TF-IDF value for the given keyword t_i as its best cluster cb_{t_i} .

$$cb_{t_i} = \arg \max_{c_x \in c_{t_i}} (\text{tfidf}_{i,x}) \quad (5)$$

Here, c_{t_i} denotes the set of clusters c_x that contain the keyword t_i . To mitigate redundancy, we combine n -gram keywords belonging to the same cluster. For example, the keywords “*unreal*,” “*engine*,” and “*unreal engine*” are associated with the same cluster with similar importance, indicating that the two individual words are typically used together. Thus, we retain the “*unreal engine*” and eliminate the two individual words “*unreal*” and “*engine*.”

5 SYSTEM DESIGN

We have developed a visual analysis system that leverages the Stable Diffusion model and our prompt recommendation model to assist users in interactive prompt engineering and text-to-image creation. In this section, we first describe the user interface of *PromptMagician*, followed by a detailed explanation of critical system components, including the multi-level image-prompt visualization and the image evaluation.

5.1 User Interface

The user interface (Figure 1) encompasses four views. The *Model Input View* (Figure 1A) allows users to input the text prompt and customize the model hyper-parameters. The *Image Browser View* (Figure 1B) visualizes the image collection, including model-generated and retrieved images as well as the recommended prompt keywords. The *Image Evaluation View* (Figure 1C) empowers users to establish aesthetic criteria for assessing and filtering out irrelevant images. Users can navigate through the *Image Browser View* and select specific images for further examination in *Local Exploration View* (Figure 1D).

Model Input View serves as the starting point of the image creation process. Users can input a prompt to describe the desired subjects and styles and configure the model hyper-parameters, including guidance scale and total generation. Once the range of the guidance scale is specified, the system randomly samples values within the range.

Image Browser View is the primary window for users to explore the image collection. Both the generated and retrieved images and the prompt keywords are co-embedded and visualized within the view. The images are positioned based on their semantic embedding (Section 4.2), and the keywords are positioned near their most related image clusters (Section 4.5). The images and keywords are presented in multi-level visualization to reduce the visual clutter [2, 22]. At the overview level, representative images for the clusters are visualized, and the others are replaced by translucent rectangles. As the users navigate to the detailed level, all images within the clusters and their corresponding prompt keywords are gradually revealed. Users can hover over an image to view its detailed information or click a keyword to highlight specific images containing it. The title bar of *Image Browser View* includes checkboxes for controlling the visibility of the image types (i.e., generated or retrieved) and keywords.

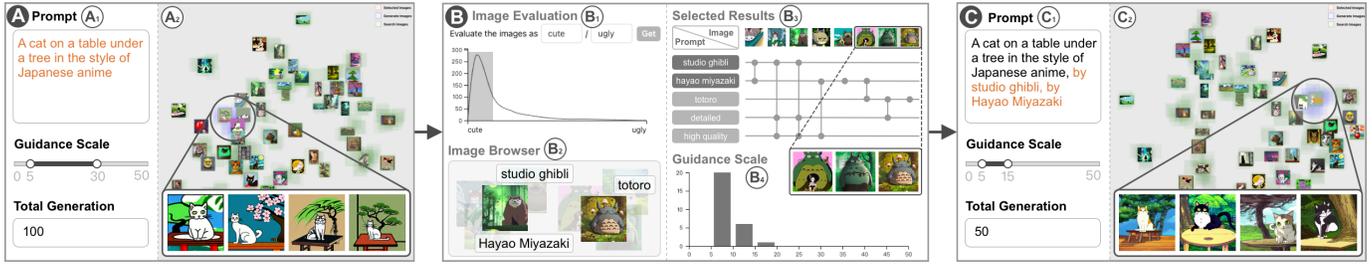


Fig. 4: The first usage scenario showcases how our system supports the user in refining the prompt for the desired image style. (A) The user tries the first prompt but is not satisfied with the model-generated images. (B) She then explores similar images and receives recommended prompt keywords that clarify the desired image style. (C) Finally, she successfully obtains the desired results on the second attempt.

Image Evaluation View allows users to specify criteria for filtering images from multiple aspects (e.g., aesthetic). Users can input two keywords (e.g., *simple* and *complex*) that represent two poles of an evaluation criterion (e.g., image complexity). The second keyword is optional and will be set as the opposite of the first keyword using “not” if not specified. The system then rates all images based on the two keywords (Section 5.3) and visualizes the rating distribution to support brush interaction for image filtering. At the top of the view, we recommend some common pairs of keywords for image evaluation [53].

Local Exploration View comprises three panels. The *Result Details Panel* presents detailed information about the selected results, enabling users to view the high-resolution images and explore their prompts and hyper-parameter values. The *Prompt Keywords Panel* not only presents the prompt keywords (sorted by the keyword importance introduced in Section 4.4) for the selected images but also visualizes their interrelationship with the images using BioFabric [25], a tabular layout for graph visualization. Each point in the table indicates that a keyword is used by an image. The users can explore how frequently the keywords are used by the selected images, or compare the images with different keywords to better understand the effect of the keywords. The users can also select the keywords to highlight them in their context in *Result Details Panel* for easy location. The *Guidance Scale Panel* visualizes the distribution of the guidance scales of the selected images in a histogram. Both the recommended prompt keywords and the range of the guidance scale can assist users in refining their original input to obtain better results from the Stable Diffusion model.

5.2 Multi-Level Image-Prompt Visualization

All generated and retrieved images are projected into a 2-D space with the t-SNE dimensional redundancy algorithm. To minimize visual clutter at the overview level and support semantic zoom [4, 27, 29, 40], we employ the hierarchical structure of image clusters to construct multi-level visualization of images. For the image clusters, we choose the images closest to the cluster center as the representative images and present them at the overview level. For the prompt keywords mapped to the image clusters, we position them near the images whose prompts contain the keywords. The positions of the keywords p_{ti} are calculated by the weighted average of their image positions:

$$p_{ti} = \frac{\sum n_{i,j} \times p_j}{\sum n_{i,j}} \quad (6)$$

Here, p_j donates the position of the images, and $n_{i,j}$ donates the number of times the keyword appears in the prompt of the images. If multiple keywords are positioned in the same position, we add a small random shift to the position of the keyword to avoid overlap. The prompt keywords are also visualized in multi-level, corresponding to the level of their respective clusters.

5.3 Image Evaluation

Image evaluation allows users to specify criteria for images from both objective aspects [41] (e.g., plane, cat, and dog) and subjective aspects (e.g., quality perception [8, 67, 72] and abstract perception [1, 30, 39, 53]). Inspired by prior work [53], we use CLIP model [41] to capture

the relationship between text and visual perception. We use pairs of opposing texts related to human perception (e.g., *real* and *abstract*) to fill in the pre-defined template for image evaluation (i.e., [text] image). We then calculate the feature cosine similarity s_i of each image with the two texts and compute the image rating (represented by \bar{s}) between the pair of keywords on a scale of $[0, 1]$ using Softmax:

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}} \quad (7)$$

Utilizing two opposing keywords transforms the image evaluation task to a binary classification, which can effectively reduce the ambiguity that arises from using a single keyword [53]. The closer \bar{s} is to 0 or 1, the closer the image is to the keyword t_1 or t_2 . Sometimes it is not easy for users to specify the second opposing keyword. We make it optional and generate the opposite of the first keyword using the negative word “not.” The image evaluation strategy can be further extended to support natural language sentences to better distinguish the nuances in the images, and we leave that for future work.

6 USAGE SCENARIOS

In this section, we present two usage scenarios that showcase the utility of our system. The first usage scenario demonstrates how our system helps the user improve her text prompt and adjust the model hyper-parameters to achieve the desired image result with the target image style. The second usage scenario exemplifies how our system inspires the user with a broad creative goal and guides him in gradually clarifying the subjects of the images and improving the quality of the generated results.

6.1 Scenario 1: Prompting for the Desired Image Style

In this scenario, the user desires to create an anime-style work featuring her cat. She provides the initial prompt “a cat on a table under a tree in the style of Japanese anime.” in *Model Input View* (Figure 4A₁). Being unfamiliar with the model hyper-parameters, she sets the guidance scale range from 5 to 30 and requires the system to generate 100 images at a time. Following a brief waiting period, the system produced a collection of generated images in the *Image Browser View* (Figure 4A₂). After browsing through the results, the user finds that the generated results do not align with her expectations.

To refine her prompts toward the desired style, the user explores similar images for inspiration. Given the considerable number of search results, the user specifies artistic criteria based on the keyword pair “cute” and “ugly” to select images (Figure 4B₁). In response, the system visualizes the distribution of all images’ ratings on this criteria. The user brushes to select the most “cute” images. From the remaining images in the *Image Browser View*, she identifies some images characterized by a Totoro style (Figure 4B₂). The presence of prompt keywords surrounding the images confirms her finding. As the user prefers this style, she navigates the detailed level of the view and selects these images using a brush interaction for deeper exploration.

The details of the selected images are listed in the *Local Exploration View*. After a brief review of the results, the user proceeds to examine the suggested prompt keywords in the *Prompt Keywords Panel*

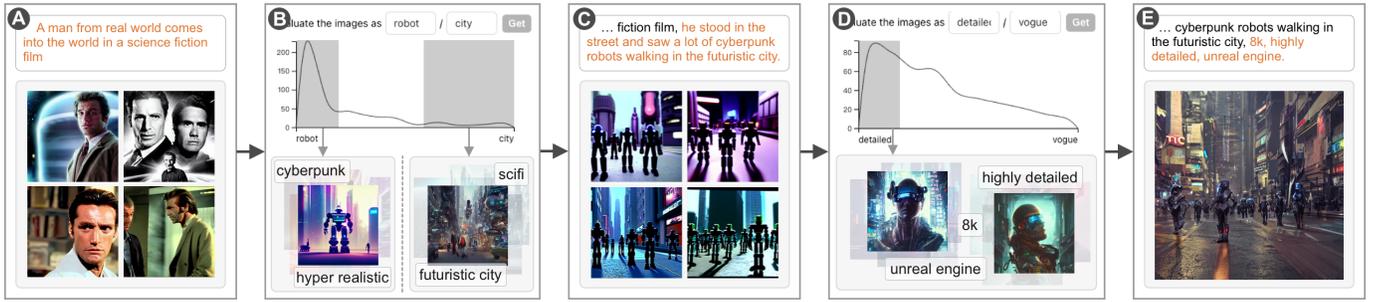


Fig. 5: The second usage scenario presents how our system facilitates open-ended creation. (A) The user inputs the first prompt and gets unexpected image results. (B) So he overviews the image collection and explores the prompt keywords for city and robot images separately. (C) Then he revises the original prompt and gets a new collection of images. (D) To improve the details of the images, the user specifies criteria to select highly detailed images. (E) With the recommended keywords, the user finally gets a satisfied image result.

(Figure 4B₃). She finds that the top-recommended prompt keywords (i.e., “Hayao Miyazaki,” “Studio Ghibli,” and “Totoro”) are strongly associated with the style and frequently used by the selected images. However, the “Totoro” keyword has a more substantial impact on the subject of the images, which is not what the user expects. Therefore, the user discards “Totoro” and chooses the other two keywords to refine her original prompt (Figure 4C₁). Additionally, the user specifies a narrower range of guidance scales (5-10), referencing the selected images (Figure 4B₄). As a result, the system generates a collection of images in the style of the Totoro movie (Figure 4C₂). Finally, the user selects a favorite image as the outcome of this creation process.

6.2 Scenario 2: Prompting for Open-Ended Creation

In this scenario, the user begins with the broad creative goal of illustrating a picture of a future world. He inputs the prompt “a man from the real world comes into the world in a science fiction film” to leverage the imaginative capabilities of the generative model (Figure 5A). However, the model-generated results primarily consist of freeze-frames of male characters from old science fiction films, which are not the intended subjects. Therefore, the user explores similar retrieved images to refine the image subjects.

Upon reviewing the overview of the *Image Browser View*, the user recognizes that the retrieved images primarily feature individuals (robots or cyborgs) or cities, inspiring the user to create images using these subjects. To explore these subjects and their corresponding keywords further, the user specifies a pair of keywords, “robot” and “city,” for image selection (Figure 5B). Using the rating distribution chart, the user then successively explores the images “close” to robots and cities through brush interaction. With the help of recommended prompt keywords, the user improves his original prompt by adding a second sentence that better clarifies the image subjects “he stood in the street and saw a lot of cyberpunk robots walking in the futuristic city.”

For the image results of the second generation (Figure 5C), the user notices that the city scene has a futuristic feel, but the robots lack sufficient detail and appear vague. Therefore, the user continues to explore the retrieved images that are “close” to robots and specify additional criteria for image quality using the keywords “detailed” and “vogue” (Figure 5D). Upon examining some detailed robot images, the user identifies a set of prompt keywords and phrases, such as “8k,” “highly detailed,” and “unreal engine.” He adds these keywords and phrases to his prompt (Figure 5E). This time, the results returned contain more images with high details and textures, from which the user selects an image as the final creation outcome.

7 USER STUDY

We conducted a user study to evaluate the effectiveness and usability of our system in facilitating interactive prompt engineering and image creation. Specifically, we aim to evaluate (1) the helpfulness of the prompting recommendation model, (2) the effectiveness and usability of the overall system, and (3) the creativity support compared with two baseline systems that mimic real-world text-to-image creation.

7.1 Participants

We recruited 12 participants (P1-P12, four females and eight males, aged 24-32) from a local university through an internal school forum. The participants are mainly undergraduate and master’s students from various disciplines, including industrial design, digital media, computer science, and literature. They had more or less experience with text-to-image generation tools but lacked sufficient knowledge about the generative models and how best to use them.

7.2 Baseline Systems

We designed two baseline systems for comparative study alongside our system. All three systems utilize the Stable Diffusion model as the backbone model for text-to-image creation.

Baseline A provides the same similar image retrieval as our system without the prompt keyword recommendation feature. This baseline mimics a typical creation scenario where users search related artwork created by previous authors on public platforms (e.g., Lexica) to gain inspiration for their own text-to-image creation.

Baseline B presents only the model-generated images in the *Image Browser View* without the image retrieval and prompt keyword recommendation. However, it introduces Promptist [15], an automatic prompting method that automatically helps users refine the prompts to improve the aesthetic quality of generated images.

7.3 Procedure and Tasks

Introduction (15 min). We first provided a brief introduction of the research background, including the research motivation and the study protocol. Next, we gathered demographic information from the participants and asked for their consent to record their operations and results for further analysis. We then introduced the feature of the system views and demonstrated their usage with clear examples [66].

Target replication training (15 min). To help users become familiar with the systems, we set a training task that requires the participants to replicate some given images (e.g., the images in Figure 5E) using our system and Promptist. Following the Jeopardy evaluation methodology [13], we did not provide textual descriptions about the target images and the metadata (e.g., guidance scales and random seeds).

Open-ended creation (60 min). In this stage, the participants were required to conduct open-ended creation tasks under a specific theme (e.g., city) without detailed constraints on the target of the creation, which aimed to evaluate the overall features and usability of the systems [7, 12, 50]. The participants were required to utilize three systems (up to 20 minutes per system) for image creation. For each system, the participants needed to (1) choose a theme for image creation, (2) write a prompt for the initial attempt, and (3) iteratively improve the prompts using the system. The order of the systems was counterbalanced.

Semi-structured interview (30 min). We asked the participants to complete a five-point Likert-scale questionnaire to evaluate the effectiveness and usability (Figure 6) of our system and the creativity support of all three systems (Figure 7). Finally, we conducted an interview with the participants to collect their feedback for further analysis.

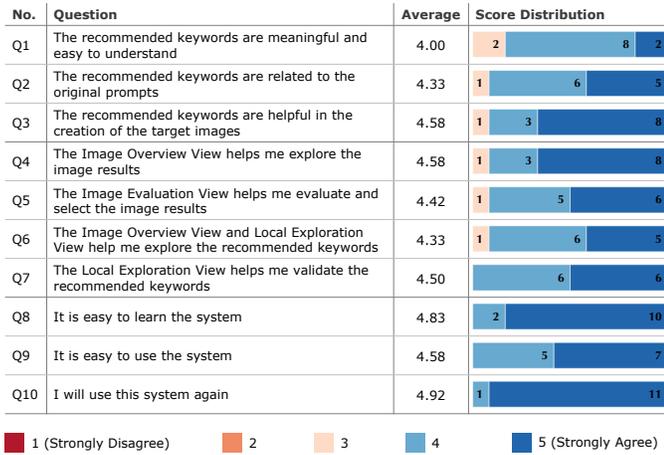


Fig. 6: The results of the questionnaire regarding the effectiveness and usability of the visual system and prompt recommendation model.

7.4 Results Analysis

All the participants completed the training and open-ended creation tasks and experienced the system’s functions. Based on the user ratings in the questionnaire and feedback received during the interview, we discuss the effectiveness and usability of the system and prompting model. We also report the difference between our system and the other two baselines. Finally, we report insights into user patterns when using our system for prompt engineering as well as areas for improvement.

7.4.1 Effectiveness of the Prompting Model

Most of the participants thought the recommended keywords were **meaningful** and **easy to understand** (Q1). P3 commented, “*I can easily understand which parts of the prompts I can improve with the keywords.*” Sometimes the participants were not familiar with some artists’ names, such as “*Leonid Afremov*” (P8). Most of the participants agreed that the recommended prompt keywords were **related** to their original prompts (Q2). P6 found that the prompt keywords might be directly related (*e.g.*, the style or composition of the images) or indirectly related (*e.g.*, commonly paired objects in similar images) to the prompts. P7 pointed out that some recommended keywords were for general purposes and were not related to specific subjects, such as “*highly detailed.*” The recommended keywords were considered to be **helpful** in the prompt improvement (Q3) regarding stylization and image quality. For instance, P10 appreciated that the recommended keywords “*summary afternoon*” helped him better control the hue of the sky in the images. P4 felt impressed that the recommended keyword “*beautiful*” significantly improved the aesthetic qualities of the images.

7.4.2 Effectiveness of the Visual System

The *Image Browser View* was appreciated by the participants for **image exploration** (Q4). The semantic-based visualization “*effectively grouped similar images together*” (P3) and facilitated users to “*batch select images for detailed comparison*” (P7). The multi-level visualization further improved the exploration experience as it “*provided natural switching between the levels*” (P4) and “*reduced cognitive load*” (P9). The *Image Evaluation View* helped most participants **evaluate and select images** (Q5) as it supported free-form criteria, which helped the users “*narrow down the exploration space and accelerate the exploration process*” (P1). It was interesting that it even worked for some special aspects, like the strength (strong or weak) of the animals in the images. The *Image Browser View* and *Local Exploration View* were found to be useful for **exploring the recommended keywords** (Q6). The participants confirmed that visualizing the keywords near the images helped them understand both. In addition, highlighting images around keywords “*showcased the impact scope of the keywords*” (P1) and “*provided hints on image selection*” (P8). For the *Prompt Keywords Panel*, the participants appreciated visualizing the correlation between

keywords and images. P1 commented that it helped discover common keyword combinations and what could be generated using them. Most participants confirmed that the *Local Exploration View* helped them **validate the recommended keywords** (Q7). P8 mentioned that comparing the images containing different keywords was quite useful to exclude keywords that were frequently used but not related to the expected aspects. The highlighting of the keywords in their original prompts also helped users “*understand the context of keywords*” (P9).

7.4.3 Usability

All the participants agreed that our system was **easy to learn** (Q8) and **easy to use** (Q9). The participants thought that the workflow of our system was intuitive and the interface was user-friendly. They also appreciated the system’s reminders of estimated times for image generation, which helped them make better trade-offs between the number of images generated and waiting time. P6 suggested adding a mini-map for the *Image Browser View* to better support navigation. Lastly, all the participants expressed their willingness to **use our system again** for creation tasks in the future (Q10). P12 stated, “*Exploring the images in the system was immersive, and I enjoyed the process.*”

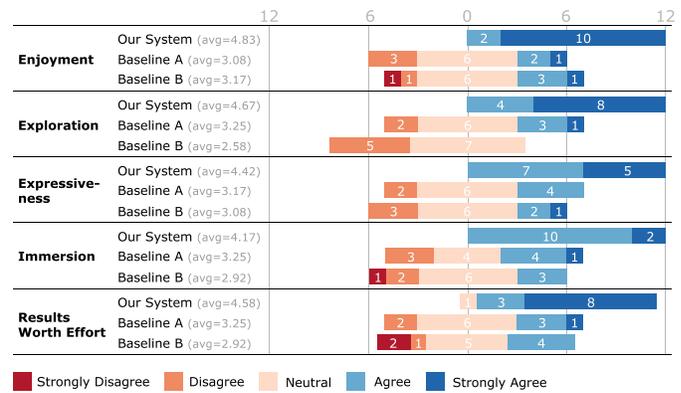


Fig. 7: The results of the questionnaire regarding the creativity support of our system and two baseline systems.

7.4.4 Creativity Support Comparison

To compare our system with two baselines in facilitating creation tasks, we asked the participants to rate three systems based on the Creativity Support Index [9]. We excluded the inapplicable “collaboration” dimension. The results are shown in Figure 7. Overall, our system outperforms the baselines in all dimensions, suggesting higher effectiveness in facilitating text-to-image creation. We have summarized user feedback on their experience as follows:

Our system vs. Baseline A. When browsing the images in Baseline A, the participants always had to review the original prompts one by one and compare them to find useful keywords. This became more tedious when comparing multiple long prompt sentences. In contrast, our system recommended important keywords for the image results and highlighted the keywords in the context, thus “*facilitating the selection and analysis of alternative keywords*” (P11) and even “*motivating more creation attempts*” (P5). Moreover, some participants preferred to select multiple keywords and copy them all at once, which helped them improve input efficiency and focus more on the creation process.

Our system vs. Baseline B. When using Baseline B for prompt improvement, the participants found that it automatically added a set of keywords to the original prompts, mostly the disciplines of art and the names of the artists. While this strategy was generally perceived as “*straightforward*” (P2), it often “*led to the emergence of unexpected image styles*” (P7) and sometimes even “*affected the accuracy of the image subjects*” (P3). Notably, grappling with the suggested textual keywords in the absence of visual hints posed a considerable challenge. Our system visualized the prompt keywords and images together, which “*facilitated users in comprehending the effect of the keywords*” (P11) and even “*stimulated greater creativity in image generation*” (P4).

7.5 Notable Observations

We have observed three prominent patterns in the way users craft their prompts. The first and most common pattern involved users starting with basic sentences and progressively enhancing the image details by adding new keywords, as described in the usage scenarios. The second pattern arose when users were unable to achieve desired results with detailed descriptions and needed summarizing suggestions. For instance, using “*summer afternoon*” to represent the coloring of the sky and clouds achieved the desired color effect and avoided color confusion with other objects. The third type involved the selective replacement of certain keywords. For instance, the prompt “*a woman in an Arabesque costume lying on a soft Persian carpet*” generated an image of a woman standing against a carpet backdrop. However, replacing “carpet” with “pillow” improved both body posture and background. These prompting patterns and user intents can be detected for adaptive recommendations to select and refine keywords.

8 DISCUSSION

Apart from the case studies and user study, we conducted interviews with five experts from the Midjourney creator community to evaluate *PromptMagician*. Their expertise spans from professional creators to advertisement designers, providing a diverse spectrum of insights. Each interview session consists of case study walkthroughs, open-ended exploration, and semi-structured interviews. Here, we distill design implications and discuss our system’s generalizability and limitations.

8.1 Design Implications

Ideating with text-to-image creation. *PromptMagician* employs a breadth-first search approach to navigate the vast artistic search space. This approach’s efficiency surpasses traditional manual drafting methods, making it particularly beneficial for those more imaginatively inclined than technically skilled. By turning abstract ideas into concrete visuals, it brings early-stage clarity to synchronize ideas across collaborators and simplify subsequent fine-tuning tasks, thus streamlining the overall creative process. Unlike conventional image browsing tools, *PromptMagician* combines querying image databases with generating ad-hoc images, harnessing the ever-evolving capabilities of generative AI. However, a knowledge gap exists for both novices and experts in adapting their prompt strategies alongside the model improvements, primarily because most users lack systematic exploratory methods for effective prompting. Our approach to prompt keyword discovery is viable for exploring the expanding capabilities of text-to-image models.

Balancing between the brevity and effectiveness of prompts. Text-to-image creation has shifted the paradigm from the mechanical task of drawing to the perceptual notion of aesthetics. While drawing more details usually enhances an artwork, our findings indicate that for crafting prompts, less is often more. Short, concept-focused prompts tend to yield better results than long, descriptive ones. For instance, adding negative prompts to rectify a three-handed human portrait might unintentionally result in a handless figure because of the linguistic ambiguity and unpredictable control of prompt keywords over specific visual elements. This phenomenon, which we term “*semantic contamination*,” discourages the unnecessary inclusion of ambiguous keywords but encourages the appropriate selection of effective keyword combinations. This suggests the necessity for further research to identify the optimal balance between brevity and effectiveness of the prompt keywords to achieve the desired outcomes.

Incorporating aesthetic aspects into machine evaluation. Experts appreciate the efficiency that *PromptMagician* introduces to the artistic creation process, as it supports machine evaluation of massive generation and database recommendation. This evaluation can extend beyond the CLIP model to include existing art-based evaluators and integrate multiple concepts to reflect the multi-dimensional nature of aesthetic values. For instance, detecting the “uncanny valley” [32] requires the evaluation of human-like objects from several perspectives (e.g., facial resemblance and limb authenticity), which opposing keywords alone may struggle to capture. Future research could explore other evaluation metrics (e.g., color analysis [11] and art appreciation theories [75]) to enhance artistic sensibilities within image evaluation.

Tracking the iterative creation process. Our interactive prompt engineering workflow supports users in articulating their requirements in the creation and bridging knowledge gaps in describing visual elements and artistic styles. However, users often struggle with when to terminate the fine-tuning process in the user study due to uncertainty about the model’s capability or inability to achieve the desired results [56]. Currently, experts rely on their experience and heuristics to address the issue, highlighting a need for guidance in the iterative process.

Fine-tuning with multimodal interactions. AI-generated artworks are not bounded by physical laws or artistic principles, which often require fine-tuning of the generative output. Image conditioning methods [33, 42, 43] support text-guided inpainting to manipulate specific areas within images. Another thread of research, exemplified by ControlNet [71] and DragGan [38], provides interactive modifications (e.g., specifying object boundaries and layouts), facilitating an iterative improvement process in text-to-image creation. Experts express keen interest in these intuitive interactive tools that blend seamlessly into their existing workflows. The combination of these research directions could lead to a more flexible and efficient creative process, especially if visualization tools can support high-level stylistic changes with textual prompts and nuanced detail adjustments through interactive inputs.

8.2 System Generalizability

Besides facilitating prompt engineering for text-to-image generation, our system can be generalized to various applications, such as content moderation and model analysis. For content moderation, our system can help examine and mitigate the potential misuse of text-to-image models. The prompt keyword recommendation and *Local Explanation View* empower content reviewers to effectively identify and scrutinize harmful content prompted by specific groups of words and limit the usage of such prompts. For model analysis, the *Image Browser View* allows researchers to evaluate a large number of generated model outputs. This feature facilitates the assessment of key aspects such as precision, stylization, and diversity in the model’s output, thereby offering a comprehensive understanding of the model’s strengths and weaknesses.

8.3 Limitations and Future Work

PromptMagician currently allows users to configure text prompts and guidance scales. Other hyper-parameters (e.g., the *inference step* for image denoising) can improve image quality but also pose challenges to the trial-and-error process. We plan to support tuning multiple hyper-parameters to achieve flexible human control without increasing the learning cost or usage complexity.

PromptMagician retrieves similar images from DiffusionDB and identifies prompt keywords related to user prompts. An interesting future work is to apply deep learning models, especially large language models, for prompt improvement. For example, GPT-4 [34] has been empowered with multi-modal data processing capability. It is possible to use GPT-4 to recommend prompt keywords for the image clusters or automatically revise user prompts according to the images of interest to the users. Integrating such models facilitates human-in-the-loop prompt improvement, leading to effective text-to-image generation.

9 CONCLUSION

This paper presents *PromptMagician*, a visual analytic system for interactive prompt engineering in text-to-image creation. The system helps the users generate and explore a collection of image results and iteratively refine the input prompt. We design a prompt recommendation model to recommend special and related prompt keywords for user prompts, which are mined from DiffusionDB with semantic-based retrieval and hierarchical keyword extraction. Both the prompt keywords and their corresponding images are co-embedded in 2D space to facilitate interactive exploration and support personalized evaluation. We present two usage scenarios of our system and conduct a user study and expert interviews to validate its effectiveness and usability. The results not only show that *PromptMagician* recommends useful prompt keywords and facilitates interactive exploration, but also provide new insights for designing and improving prompting methods and the visual system for text-to-image creation.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful comments. This paper is supported by the National Natural Science Foundation of China (62132017) and the Fundamental Research Funds for the Central Universities (226-2022-00235).

REFERENCES

- [1] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas. Artemis: Affective language for visual art. In *Proc. CVPR*, pp. 11569–11579. IEEE/CVF, Piscataway, 2021. doi: 10.1109/CVPR46437.2021.01140 6
- [2] D. Bertucci et al. Dendromap: Visual exploration of large-scale image datasets for machine learning with treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):320–330, 2022. doi: 10.1109/TVCG.2022.3209425 2, 5
- [3] 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. In *Proc. NeurIPS*, vol. 33, pp. 1877–1901, 2020. 2, 3
- [4] T. Buring, J. Gerken, and H. Reiterer. User interaction with scatterplots on small screens - A comparative evaluation of geometric-semantic zoom and fisheye distortion. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):829–836, 2006. doi: 10.1109/TVCG.2006.187 6
- [5] P. J. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv*, 2022. doi: 10.48550/arXiv.2210.04133 1
- [6] C. Chen, J. Wu, X. Wang, S. Xiang, S.-H. Zhang, Q. Tang, and S. Liu. Towards better caption supervision for object detection. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1941–1954, 2022. doi: 10.1109/TVCG.2021.3138933 2
- [7] Z. Chen and H. Xia. Crossdata: Leveraging text-data connections for authoring data documents. In *Proc. CHI*, pp. 95:1–95:15. ACM, NY, 2022. doi: 10.1145/3491102.3517485 7
- [8] M. Cheon, S. Yoon, B. Kang, and J. Lee. Perceptual image quality assessment with transformers. In *Proc. CVPR*, pp. 433–442. IEEE/CVF, Piscataway, 2021. doi: 10.1109/CVPRW53098.2021.00054 6
- [9] E. Cherry and C. Latulipe. Quantifying the creativity support of digital tools through the creativity support index. *ACM TOCHI*, 21(4):1–25, 2014. doi: 10.1145/2617588 8
- [10] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186. ACL, Stroudsburg, 2019. doi: 10.18653/v1/n19-1423 3
- [11] Y. Feng, J. Chen, K. Huang, J. K. Wong, H. Ye, W. Zhang, R. Zhu, X. Luo, and W. Chen. iPoet: interactive painting poetry creation with visual multimodal analysis. *Journal of Visualization*, 25(3):671–685, Jun 2022. doi: 10.1007/s12650-021-00780-0 9
- [12] Y. Feng, X. Wang, B. Pan, K. K. Wong, Y. Ren, S. Liu, Z. Yan, Y. Ma, H. Qu, and W. Chen. XNLI: Explaining and diagnosing nli-based visual data analysis. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–14, 2023. doi: 10.1109/TVCG.2023.3240003 7
- [13] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proc. UIST*, pp. 489–500. ACM, NY, 2015. doi: 10.1145/2807442.2807478 7
- [14] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. In *Proc. ACL/IJCNLP*, pp. 3816–3830. ACL, Stroudsburg, 2021. doi: 10.18653/v1/2021.acl-long.295 2
- [15] Y. Hao, Z. Chi, L. Dong, and F. Wei. Optimizing prompts for text-to-image generation. *arXiv*, 2022. doi: 10.48550/arXiv.2212.09611 7
- [16] J. He, X. Wang, K. K. Wong, X. Huang, C. Chen, Z. Chen, F. Wang, M. Zhu, and H. Qu. Videopro: A visual analytics approach for interactive video programming. *arXiv*, 2023. doi: 10.48550/arXiv.2308.00401 2
- [17] J. Ho et al. Imagen video: High definition video generation with diffusion models. *arXiv*, 2022. doi: 10.48550/arXiv.2210.02303 1
- [18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, vol. 33, pp. 6840–6851, 2020. doi: 10.48550/arXiv.2006.11239 3
- [19] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How can we know what language models know? *Transactions of the ACL*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324 2
- [20] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proc. EMNLP*, pp. 3045–3059. ACL, Stroudsburg, 2021. doi: 10.18653/v1/2021.emnlp-main.243 2
- [21] P. P. Liang, Y. Lyu, G. Chhablani, N. Jain, Z. Deng, X. Wang, L.-P. Morency, and R. Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models. In *Proc. ICLR*, 2022. 2
- [22] Y. Lin, K. Wong, Y. Wang, R. Zhang, B. Dong, H. Qu, and Q. Zheng. Taxthemis: Interactive mining and exploration of suspicious tax evasion groups. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):849–859, 2021. doi: 10.1109/TVCG.2020.3030370 5
- [23] V. Liu and L. B. Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proc. CHI*, pp. 1–23, 2022. doi: 10.1145/3491102.3501825 1, 2, 4
- [24] V. Liu, H. Qiao, and L. Chilton. Opal: Multimodal image generation for news illustration. In *Proc. UIST*, pp. 384:1–384:23. ACM, NY, 2022. doi: 10.1145/3526113.3545621 2
- [25] W. J. Longabaugh. Combing the hairball with biofabric: a new approach for visualization of large networks. *BMC bioinformatics*, 13(1):1–16, 2012. doi: 10.1186/1471-2105-13-275 6
- [26] E. Loper and S. Bird. Nltk: The natural language toolkit. *arXiv*, cs.CL/0205028, 2002. doi: 10.48550/arXiv.cs/0205028 5
- [27] F. D. Luca, M. I. Hossain, S. G. Kobourov, and K. Börner. Multi-level tree based approach for interactive graph visualization with semantic zoom. *arXiv*, 2019. doi: 10.48550/arXiv.1906.05996 6
- [28] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Proc. NeurIPS*, 30:4765–4774, 2017. 2
- [29] S. L’Yi, Q. Wang, F. Lekschas, and N. Gehlenborg. Gosling: A grammar-based toolkit for scalable and interactive genomics data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):140–150, 2022. doi: 10.1109/TVCG.2021.3114876 6
- [30] C. Ma, C. Yang, X. Yang, and M. Yang. Learning a no-reference quality metric for single-image super-resolution. *Comput. Vis. Image Underst.*, 158:1–16, 2017. doi: 10.1016/j.cviu.2016.12.009 6
- [31] J. MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297. University of California Los Angeles LA USA, 1967. 4
- [32] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. doi: 10.1109/MRA.2012.2192811 9
- [33] A. Q. 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. In *Proc. ICML*, vol. 162, pp. 16784–16804. PMLR, 2022. doi: 10.48550/arXiv.2112.10741 1, 2, 3, 9
- [34] OpenAI. GPT-4 technical report. *arXiv*, pp. 2303–08774, 2023. doi: 10.48550/arXiv.2303.08774 9
- [35] J. Oppenlaender. Prompt engineering for text-based generative art. *arXiv*, 2022. doi: 10.48550/arXiv.2204.13988 2
- [36] J. Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *arXiv*, 2022. doi: 10.48550/arXiv.2204.13988 2, 4
- [37] L. Ouyang et al. Training language models to follow instructions with human feedback. *arXiv*, 2022. doi: 10.48550/arXiv.2203.02155 2
- [38] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *Proc. SIGGRAPH*. ACM, NY, 2023. doi: 10.48550/arXiv.2305.10973 9
- [39] R. Panda, J. Zhang, H. Li, J. Lee, X. Lu, and A. K. Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proc. ECCV*, pp. 594–612. Springer, Berlin, 2018. doi: 10.1007/978-3-030-01216-8_36 6
- [40] K. Perlin and D. Fox. Pad: an alternative approach to the computer interface. In *Proc. SIGGRAPH*, pp. 57–64. ACM, NY, 1993. doi: 10.1145/166117.166125 6
- [41] A. Radford et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, vol. 139, pp. 8748–8763. PMLR, 2021. doi: 10.48550/arXiv.2103.00020 2, 4, 6
- [42] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022. doi: 10.48550/arXiv.2204.06125 1, 2, 3, 9
- [43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pp. 10684–10695. IEEE, Piscataway, 2022. doi: 10.1109/CVPR52688.2022.01042 1, 2, 3, 9

- [44] T. L. Scao and A. M. Rush. How many data points is a prompt worth? In *Proc. NAACL-HLT*, pp. 2627–2636. ACL, Stroudsburg, 2021. doi: 10.18653/v1/2021.naacl-main.208 1
- [45] T. Schick and H. Schütze. It’s not just size that matters: Small language models are also few-shot learners. In *Proc. NAACL-HLT*, pp. 2339–2352. ACL, Stroudsburg, 2021. doi: 10.18653/v1/2021.naacl-main.185 2
- [46] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proc. EMNLP*, pp. 4222–4235. ACL, Stroudsburg, 2020. doi: 10.18653/v1/2020.emnlp-main.346 2
- [47] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, vol. 37, pp. 2256–2265. PMLR, JMLR.org, 2015. doi: 10.48550/arXiv.1503.03585 3
- [48] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *Proc. ICLR*. OpenReview.net, 2021. doi: 10.48550/arXiv.2010.02502 3
- [49] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. doi: 10.1108/00220410410560573 2
- [50] A. Srinivasan and J. Stasko. How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE CG&A*, 40(4):96–103, 2020. doi: 10.1109/MCG.2020.2986902 7
- [51] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1146–1156, 2022. doi: 10.1109/TVCG.2022.3209479 2
- [52] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4
- [53] J. Wang, K. C. Chan, and C. C. Loy. Exploring clip for assessing the look and feel of images. In *Proc. AAAI*. AAAI Press, Menlo Park, CA, 2022. doi: 10.48550/arXiv.2207.12396 6
- [54] X. Wang, F. Cheng, Y. Wang, K. Xu, J. Long, H. Lu, and H. Qu. Interactive data analysis with next-step natural language query recommendation. *arXiv*, 2022. doi: 10.48550/arXiv.2201.04868 2
- [55] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2021. doi: 10.1109/TVCG.2021.3114794 2
- [56] X. Wang, Z. Wu, W. Huang, Y. Wei, Z. Huang, M. Xu, and W. Chen. Vis+ ai: integrating visualization with artificial intelligence for efficient data analysis. *Frontiers of Computer Science*, 17(6):1–12, 2023. doi: 10.1007/s11704-023-2691-y 9
- [57] Y. Wang, S. Shen, and B. Y. Lim. Reprompt: Automatic prompt editing to refine ai-generative art towards precise expressions. *arXiv*, 2023. doi: 10.48550/arXiv.2302.09466 2
- [58] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv*, 2022. doi: 10.48550/arXiv.2210.14896 1, 2, 3, 4, 5
- [59] L. Weng, M. Zhu, K. K. Wong, S. Liu, J. Sun, H. Zhu, D. Han, and W. Chen. Towards an understanding and explanation for mixed-initiative artificial scientific text detection. *arXiv*, 2023. doi: 10.48550/arXiv.2304.05011 2
- [60] K. K. Wong, X. Wang, Y. Wang, J. He, R. Zhang, and H. Qu. Anchorage: Visual analysis of satisfaction in customer service videos via anchor events. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2023. doi: 10.1109/TVCG.2023.3245609 2
- [61] T. Wu, E. Jiang, A. Donsbach, J. Gray, A. Molina, M. Terry, and C. J. Cai. Promptchainer: Chaining large language model prompts through visual programming. In *Proc. CHI*, pp. 1–10. ACM, NY, 2022. doi: 10.1145/3491101.3519729 2
- [62] T. Wu, M. Terry, and C. J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proc. CHI*, pp. 385:1–385:22. ACM, NY, 2022. doi: 10.1145/3491102.3517582 2
- [63] J. Xia, L. Huang, W. Lin, X. Zhao, J. Wu, Y. Chen, Y. Zhao, and W. Chen. Interactive visual cluster analysis by contrastive dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):734–744, 2022. doi: 10.1109/TVCG.2022.3209423 2
- [64] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu. A semantic-based method for visualizing large image collections. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2362–2377, 2018. doi: 10.1109/TVCG.2018.2835485 2, 5
- [65] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. Ward. Semantic image browser: Bridging information visualization with automated intelligent image analysis. In *Proc. VAST*, pp. 191–198. IEEE Computer Society, Los Alamitos, 2006. doi: 10.1109/VAST.2006.261425 2
- [66] L. Yang, C. Xiong, J. K. Wong, A. Wu, and H. Qu. Explaining with examples: Lessons learned from crowdsourced introductory description of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 29(3):1638–1650, 2023. doi: 10.1109/TVCG.2021.3128157 7
- [67] P. Ye, J. Kumar, L. Kang, and D. S. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proc. CVPR*, pp. 1098–1105. IEEE/CVF, Piscataway, 2012. doi: 10.1109/CVPR.2012.6247789 6
- [68] H. Zeng, X. Wang, Y. Wang, A. Wu, T.-C. Pong, and H. Qu. Gesturelens: Visual analysis of gestures in presentation videos. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi: 10.1109/TVCG.2022.3169175 2
- [69] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu. Emoco: Visual analysis of emotion coherence in presentation videos. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):927–937, 2019. doi: 10.1109/TVCG.2019.2934656 2
- [70] H. Zhang et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. ICCV*. IEEE, Piscataway, 2017. doi: 10.1109/ICCV.2017.629 3
- [71] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv*, 2023. doi: 10.48550/arXiv.2302.05543 9
- [72] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730 6
- [73] W. Zhang, J. K. Wong, Y. Chen, A. Jia, L. Wang, J.-W. Zhang, L. Cheng, and W. Chen. Scrolltimes: Tracing the provenance of paintings as a window into history. *arXiv*, 2023. doi: 10.48550/arXiv.2306.08834 2
- [74] W. Zhang, J. K. Wong, X. Wang, Y. Gong, R. Zhu, K. Liu, Z. Yan, S. Tan, H. Qu, S. Chen, and W. Chen. Cohortva: A visual analytic system for interactive exploration of cohorts based on historical data. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):756–766, 2023. doi: 10.1109/TVCG.2022.3209483 5
- [75] W. Zhang, J.-W. Zhang, K. K. Wong, Y. Wang, Y. Feng, L. Wang, and W. Chen. Computational approaches for traditional chinese painting: From the “six principles of painting” perspective. *arXiv*, 2023. doi: 10.48550/arXiv.2307.14227 9
- [76] J. Zhou, X. Wang, J. K. Wong, H. Wang, Z. Wang, X. Yang, X. Yan, H. Feng, H. Qu, H. Ying, and W. Chen. Dpviscreator: Incorporating pattern constraints to privacy-preserving visualizations via differential privacy. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):809–819, 2023. doi: 10.1109/TVCG.2022.3209391 3
- [77] H. Zhu, M. Zhu, Y. Feng, D. Cai, Y. Hu, S. Wu, X. Wu, and W. Chen. Visualizing large-scale high-dimensional data via hierarchical embedding of knn graphs. *Visual Informatics*, 5(2):51–59, 2021. doi: 10.1016/j.visinf.2021.06.002 2
- [78] M. Zhu, P. Pan, W. Chen, and Y. Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *Proc. CVPR*, pp. 5802–5810. IEEE/CVF, Piscataway, 2019. doi: 10.1109/CVPR.2019.00595 1, 3