Capability-aware Prompt Reformulation Learning for Text-to-Image Generation

Jingtao Zhan
jingtaozhan@gmail.com
Department of Computer Science and Technology, Tsinghua University
Quan Cheng Laboratory
Beijing 100084, China

Qingyao Ai*
aiqy@tsinghua.edu.cn
Quan Cheng Laboratory
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China

Yiqun Liu
yiqunliu@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Zhongguancun Laboratory
Beijing 100084, China

Jia Chen
chenjia2@xiaohongshu.com
Xiaohongshu Inc
Beijing, China

Shaoping Ma
msp@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Zhongguancun Laboratory
Beijing 100084, China

ABSTRACT

Text-to-image generation systems have emerged as revolutionary tools in the realm of artistic creation, offering unprecedented ease in transforming textual prompts into visual art. However, the efficacy of these systems is intricately linked to the quality of user-provided prompts, which often poses a challenge to users unfamiliar with prompt crafting. This paper addresses this challenge by leveraging user reformulation data from interaction logs to develop an automatic prompt reformulation model. Our in-depth analysis of these logs reveals that user prompt reformulation is heavily dependent on the individual user's capability, resulting in significant variance in the quality of reformulation pairs. To effectively use this data for training, we introduce the Capability-aware Prompt Reformulation (CAPR) framework. CAPR innovatively integrates user capability into the reformulation process through two key components: the Conditional Reformulation Model (CRM) and Configurable Capability Features (CCF). CRM reformulates prompts according to a specified user capability, as represented by CCF. The CCF, in turn, offers the flexibility to tune and guide the CRM's behavior. This enables CAPR to effectively learn diverse reformulation strategies across various user capacities and to simulate high-capability user reformulation during inference. Extensive experiments on standard text-to-image generation benchmarks showcase CAPR's superior performance over existing baselines and its remarkable robustness on unseen systems. Furthermore, comprehensive analyses validate the effectiveness of different components. CAPR can facilitate user-friendly interaction with text-to-image systems and make advanced artistic creation more achievable for a broader range of users.

CCS CONCEPTS

• **Information systems → Multimedia content creation**; **Multimedia information systems**.

KEYWORDS

prompt reformulation, text-to-image generation, log analysis

ACM Reference Format:
Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jia Chen, and Shaoping Ma. 2024. Capability-aware Prompt Reformulation Learning for Text-to-Image Generation. In *Proceedings of The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR' 24)*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the realm of intelligent information systems, effective communication between users and systems is important. Traditionally, this interaction has been facilitated through queries in search engines, serving as concise yet powerful instructions to retrieve relevant information [10, 44]. With the advent of Artificial Intelligence Generated Content (AIGC) systems like Midjourney [32], these instructions have evolved into prompts, a critical element in shaping the quality and relevance of system responses [35, 36]. Despite their importance, most users struggle to craft optimal queries or prompts [48, 50], making automatic reformulation techniques an essential component for enhancing system performance [6, 21, 34]. Similarly, in the domain of search engines, techniques such as query auto-completion [8], expansion [9], and suggestion [45] have significantly improved user experience and system efficacy, becoming indispensable features of commercial search engines [5].

While the benefits of query reformulation are well-established within search engines [3, 5, 16, 24], the exploration of prompt reformulation for AIGC systems, particularly text-to-image generation systems, is relatively limited. Text-to-image generation systems

[*] Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR' 24, July 14 - 18, 2024, Washington D.C., USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

have revolutionized the field of artistic creation, simplifying the process to unprecedented ease [23, 25–27, 29, 31, 54]. They operate by converting user-provided text prompts into visual imagery. However, their efficacy heavily depends on the quality of the input prompts. Effective prompts should conform to a specific format, precisely describes the scene, and consist of professional terminologies such as artist names [6, 36, 48]. This level of complexity in prompt writing is challenging, making learning and practice necessary for users [15, 30]. Users often rely on studying exemplary prompts shared within the community [1, 2], consulting guides on effective prompting [35, 36], and engaging in trial-and-error to refine their skills [50]. This high learning cost and constant need to reformulate prompts substantially affect the user experience.

In this paper, we focus on leveraging user-generated reformulation data from interaction logs to develop an automatic prompt reformulation model. This is motivated by the observation that users dedicate considerable effort to prompt reformulation [50], creating rich data in the interaction logs. By designing a model that builds upon these user efforts, we aim to significantly reduce the burden of manually reformulating prompts and substantially improve the user experience.

Our initial analysis of the user interaction logs reveals a significant distinction between query reformulation and prompt reformulation in text-to-image generation scenarios. Unlike query reformulation, where users benefit significantly from search results to reformulate their queries [11, 12, 19], the effectiveness of prompt reformulation for text-to-image systems relies heavily on the individual user's capability, rather than feedback from the system. Such user capability, which varies widely among individuals and generally remains consistent within a single session, leads to a wide spectrum of prompt quality and predominantly marginal reformulation improvements. For example, the initial prompts of some users may substantially surpass the reformulated prompts of others, and instances of poorly crafted initial prompts being significantly improved through reformulation are remarkably rare. This scenario contrasts sharply with query reformulation scenarios, where users often successfully find the relevant information by the end of a session [12, 16]. Consequently, unlike previous query reformulation studies that often disregard user capability in model design, we aim to design a novel framework to introduce the crucial influence of user capability in the process of prompt reformulation.

To address this challenge, we propose the Capability-aware Prompt Reformulation framework (CAPR). CAPR incorporates user capability into the reformulation process, thereby enabling effective training with user-generated data from interaction logs. It consists of two foundational components: the Conditional Reformulation Model (CRM) and Configurable Capability Features (CCF). CRM is adept at tailoring prompt reformulation according to a specified user capability, as represented by CCF. The CCF, in turn, offers the flexibility to tune and guide the CRM's behavior. The two components offer two key benefits: (1) CRM, by adapting to user capability, aligns closely with the nature of interaction log data and thus can harvest a wealth of prompt reformulation skills across different user capability levels. (2) CCF, by introducing scrutable generation capability features, allows us to control the quality of inference and generate high-quality prompts. Consequently, CRM is empowered to surpass the average user capabilities in the training data, thereby yielding superior reformulation outcomes in practical applications[1].

We conduct comprehensive experiments on standard text-to-image generation benchmarks to examine the efficacy of CAPR. Results suggest that CAPR substantially outperforms a variety of baselines, including generic language models like GPT4 and various reformulation models. Its effectiveness can also generalize to an unseen text-to-image generation system, demonstrating its robustness. A detailed ablation study also shows that CPR can generate target images based on the specified capability conditions.

In summary, our contributions are in three folds:

- To the best of our knowledge, this is the first study that utilizes interaction logs to train a prompt reformulation model for text-to-image generation.
- We provide a comprehensive analysis that differentiates prompt reformulation in text-to-image generation from traditional query reformulation in search engines
- Inspired by our analysis, we propose a novel prompt reformulation model tailored for training on prompt reformulation logs. Results demonstrate that it achieves state-of-the-art reformulation performance on standard benchmarks.

2 RELATED WORK

2.1 Text-to-Image Generation

Text-to-image generation is a rapidly evolving field in artificial intelligence that focuses on creating visual images from textual descriptions. This technology has gained considerable attention, particularly in the realm of digital art creation, as exemplified by systems like Midjourney [32] and DALLE [34]. Existing text-to-image systems mostly adopt Diffusion as the model architecture [23, 27], which can progressively transform a random noise into a coherent image with a text as guidance. The training of these diffusion models relies heavily on extensive datasets comprising images coupled with descriptive captions [42]. During training, the model learns to correlate textual descriptions with visual features, enabling it to generate relevant images for a given text input. A noteworthy aspect of this training process is that high-quality web images are usually accompanied by professional-level captions consisting of artist names and photography terminologies. Consequently, the trained models tend to favor such prompts [18, 48], which, unfortunately, are difficult to write for average users.

2.2 Query Reformulation

Query reformulation stands as an important technique in enhancing user interaction with search engines. It addresses common issues where initial queries fail to search relevant results [5]. Techniques like auto-completion and query suggestion play a crucial role in assisting users to refine their queries, thereby enhancing the likelihood of retrieving relevant information. Prior query reformulation analyses have shown that feedback from search systems helps provide relevant terms and plays a significant role in aiding query reformulation [11, 12]. Based on this, many query reformulation methods leverage terms from search results to modify the initial queries [19, 40]. Moreover, some research focuses on developing

[1] Code is open-sourced at https://github.com/jingtaozhan/PromptReformulate

reformulation models based on query logs [12, 16]. This approach typically assumes that users successfully find relevant information by the end of their session. Thus, it constructs training pairs by treating the final query in a session as a target label for training. However, as discussed in Section 4, the assumptions about the system's feedback and user reformulation data in general web search scenarios do not hold for text-to-image generation. Therefore, a tailored methodology for prompt reformulation is necessary.

2.3 Text-to-Image Prompts

Prompt quality is a crucial factor in the effectiveness of text-to-image generation systems [6, 21, 34]. Crafting a high-quality prompt, however, poses a significant challenge. It not only demands rich art knowledge like artist names and style elements but also typically requires a time-consuming iterative process of tuning and refinement [6, 36, 48]. To aid users in this endeavor, various online platforms have emerged, offering spaces for sharing well-crafted prompts [1, 2]. In addition, comprehensive guides and textbooks have been written to teach prompt crafting techniques [35, 36]. There are even marketplaces dedicated to trading high-quality prompts [43]. However, these resources often come with substantial demands of either time investment or financial cost. To alleviate the complexity of prompt crafting, it is important to develop an automated model that is capable of reformulating subpar prompts into well-crafted ones. The primary obstacle in developing such a model is the difficulty in annotating prompt reformulation pairs for training, which requires annotators with extensive experience, leading to high costs and complexities in labeling. Previous researchers bypass this obstacle by constructing synthetic refinement data [21]. They crawl high-quality prompt demonstrations from the internet and rephrase them to user languages. However, the quality of such synthesized data is subpar, which harms the performance of the trained model. Considering this issue, our approach aims to extract reformulation data from readily available interaction logs, offering a more direct and practical solution.

3 PROBLEM FORMULATION

This section formulates the core challenge for prompt reformulation in text-to-image generation scenarios.

Text-to-Image generation systems [23, 25, 27], such as Midjourney and DALL-E, represent cutting-edge intelligent tools for artistic creation. These systems work by transforming textual descriptions, known as prompts, into visual imagery [26, 29, 31, 54]. At each interaction round, the user provides a prompt that describes their envisioned image. The system interprets this text to generate a corresponding image. While these systems liberate users from the technicalities of traditional art creation, they demand high-quality prompts to accurately generate users' envisioned images. Mathematically, let p be the prompt and i be the rendered image. The text-to-image generation system is denoted as $\mathcal{G}$, with $\mathcal{G}(i|p)$ signifying the probability of generating image i from prompt p.

Evaluating the generation quality has been extensively studied. The ideal approach involves human annotators, but without a large team and comprehensive guidelines, human's diverse preferences can lead to inconsistent assessments [13, 49]. To avoid this, prior research has developed automatic evaluation models to simulate human preferences by training them on extensive human annotations [41, 49, 51]. In this study, we employ these automatic scoring models for evaluation. We denote the scoring model as f, where $f(p, i)$ indicates the likelihood of user satisfaction with the generated image i for prompt p. Consequently, the generation quality for a prompt p can be quantified as $\mathbb{E}[\mathcal{G}(i|p) \cdot f(p, i)]$.

Notion	Text-to-Image Generation	Search Scenario
p	Prompt	Query
i	Rendered image	Search result page
$\mathcal{G}$	Text-to-image system	Search engine
f	Satisfaction with image	Relevance of search results
Ω	Prompt reformulation model	Query reformulation model

Table 1: Summary of the notions for text-to-image generation and Search Engine Scenarios.

Text-to-image generation systems are usually sensitive to input prompts, making prompt crafting a form of "art" [6, 36, 48], a skill that many users do not possess. Automatic prompt reformulation stands as a critical solution to this challenge. A reformulation model acts as an intermediary, transforming an initial user prompt into a version better suited for the text-to-image system. For example, the reformulation model can refine vague descriptions and enrich the prompt with artist references. We use Ω to represent a reformulation model, and $\Omega(\hat{p}|p)$ is the probability of reformulating p to $\hat{p}$. The reformulation model is expected to maximize the generation quality, which is formulated as:

$$\max_{\Omega} \mathbb{E}[\Omega(\hat{p}|p) \cdot \mathcal{G}(\hat{i}|\hat{p}) \cdot f(p, \hat{i})] \tag{1}$$

The reformulated prompt $\hat{p}$ serves as an intermediate in the generation process. The evaluation scores are computed based on the initial prompt and the images. The reformulated prompt aims to help render better images.

To clarify the introduced notions, we summarize them in Table 1. The table also shows the corresponding interpretations in search engine scenarios, which will be elaborated in the next section where we compare prompt reformulation with query reformulation.

4 ANALYSIS OF PROMPT REFORMULATION

This section deeply analyzes how users reformulate prompts, which serves as a crucial insight for training a reformulation model on interaction logs. We first compare prompt reformulation with query reformulation. Analysis results reveal that prompt reformulation is influenced by user capability to a larger extent. Then, we validate this observation through an examination of reformulation behaviors in large-scale interaction logs.

4.1 Reformulation: Prompt vs. Query

Prompt and query reformulation share structural similarities, which facilitates examining prompt reformulation through the lens of established findings in query reformulation. As depicted in Table 1, the components in prompt and query reformulation can correspond to each other. For instance, the rendered image corresponds to a

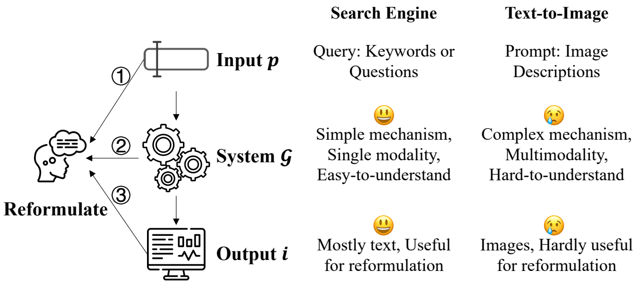

Search Engine
Text-to-Image
Input p
Query: Keywords or Questions
Prompt: Image Descriptions
①
②
System $\mathcal{G}$
Simple mechanism, Single modality, Easy-to-understand
Complex mechanism, Multimodality, Hard-to-understand
③
Reformulate
Output i
Mostly text, Useful for reformulation
Images, Hardly useful for reformulation

Figure 1: Comparing prompt reformulation with query reformulation in terms of three key factors: ① the initial input ② user’s understanding of the system’s mechanics ③ the previous system’s output. The latter two can hardly help users reformulate better prompts, indicating that prompt reformulation is a more challenging task for users.

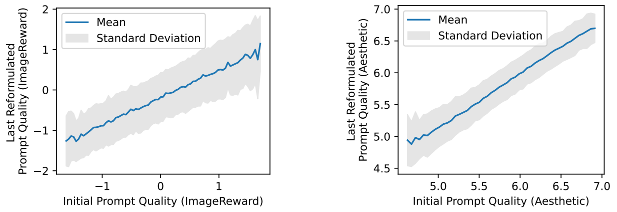
search result page, and the generation model parallels a search engine. In both contexts, users describe an information need through textual input, guiding the system to produce the desired output.

In query reformulation, a well-established body of research identifies three pivotal factors that facilitate the process [11, 12, 16, 19]: the initial query, the user’s understanding towards search engine mechanism, and the search result page. These factors are foundational in guiding effective query reformulation. The initial query lays the groundwork, providing a basis for subsequent refinements. Knowledge of the search engine’s mechanism enables users to understand how their queries are processed and related to the results. This knowledge helps them to effectively optimize the queries further. Moreover, the search results themselves often provide precise terminology and relevant context, assisting users in refining their queries with greater specificity [17, 46].

Conversely, in the realm of text-to-image generation, the last two facilitative factors are absent, making prompt reformulation a substantially more challenging task. We compare prompt reformulation with query reformulation in Figure 1. Unlike search engines, whose operating mechanisms (keyword matching, single modality) are obvious and generally understood by users, text-to-image systems often operate as “magical black boxes”. The intricate neural computations and the multimodality nature are typically beyond the user’s comprehension. Furthermore, the output of these systems, being visual imagery, does not offer textual cues, e.g., artist names or stylistic terminologies, that users can directly add to their prompts. The lack of informative feedback from the output, combined with the black-box nature of the generation systems, places a significant burden on users. They have to rely heavily on their inherent capability to intuit and imagine how different textual inputs might influence the visual output. This is a process that is less guided and more speculative compared to the more systematic and feedback-oriented process of query reformulation.

4.2 Investigation of Prompt Session Log

To validate the above analysis, we investigate user reformulation behaviour from an interaction log. The results demonstrate that prompt reformulation is indeed a challenging task that heavily depends on users’ inherent capabilities. We first introduce the dataset, our analysis methodology, and then elaborate on our findings.



(a) Overall Generation Quality measured by ImageReward [51]

(b) Aesthetic Generation Quality measured by Aesthetic Predictor

Figure 2: Comparison of generation quality between initial and reformulated prompts within a session, evaluated by ImageReward [51] and Aesthetic scoring models [41]. Results reveal limited quality improvement through users’ reformulation, suggesting that prompt quality largely depends on the user’s initial capabilities. Session contexts such as generation feedback usually offer limited assistance.

4.2.1 Dataset. In this analysis, we utilize DiffusionDB [47], a comprehensive log capturing 1.8 million interactions from 10 thousand users. This dataset includes prompts submitted by users, images generated by the system, user IDs, and timestamps. Its extensive scale and diversity enable a robust analysis of real user reformulation behavior. Since the dataset does not split interactions into sessions, we construct sessions based on timestamps and prompt topics. Specifically, inspired by the construction of search sessions, adjacent prompts that are submitted by the same user within 20 minutes and surpass a similarity threshold of 0.1 (as determined by the CLIP model [38]) are classified into the same session. We have manually examined several session splits constructed through this way. The quality of most sampled sessions is reasonable and reliable. Eventually, we obtain 30k sessions in total.

4.2.2 Analysis Methodology. We use the first prompt of each session to represent the user’s original intent and investigate how the last prompt of a session improves the generation quality. We use scoring models, namely ImageReward [51] and Aesthetic Predictor [41], to assess the quality of images generated from both the initial and the last prompts. ImageReward considers relevance and aesthetic quality, reflecting overall user satisfaction, while Aesthetic Predictor focuses solely on the visual appeal of the generated images. The results are depicted in Figures 2a and 2b.

4.2.3 Empirical Findings. The empirical results, as illustrated in Figure 2, show that the quality of user reformulation exhibits significant variance and that most users face challenges in substantially improving their initial prompts in the sessions. This is primarily attributed to the factors discussed earlier: the limited assistance from system feedback in prompt reformulation and the substantial dependence on the user’s inherent prompt-writing capability. Notably, this capability tends to remain static within a single session,

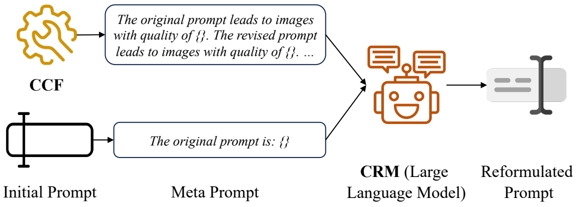

The original prompt leads to images with quality of {}. The revised prompt leads to images with quality of {}. ...
CCF
The original prompt is: {}
Initial Prompt
Meta Prompt
CRM (Large Language Model)
Reformulated Prompt

Figure 3: Architecture of Capability-aware Prompt Reformulation (CAPR). It consists of two components: the Conditional Reformulation Model (CRM) and Configurable Capability Features (CCF). Given a certain user capability indicated by CCF, CRM reformulates prompts accordingly.

leading to a scarcity of cases where users successfully reformulate poorly crafted initial prompts to high-quality ones.

Given this observed variability in reformulation quality and the prevalence of suboptimal reformulation pairs, the conventional approach of training reformulation models based on such user data presents significant challenges. Traditional query reformulation studies typically employ a direct sequence-to-sequence translation model based on reformulation pairs [14, 16, 20], but such a method may not work well in our context. Training a model on these inconsistent and suboptimal pairs could lead to substantial confusion and eventually diminish the model's overall effectiveness. If we filter the dataset to include only optimal reformulation pairs, the final size of the training data would be too small to ensure robust model training. Thus, new methods are needed to address these challenges in text-to-image prompt reformulation.

5 METHODOLOGY

Inspired by the insights gained from our previous analysis, we propose the Capability-aware Prompt Reformulation (CAPR) framework, a novel approach to effectively train a prompt reformulation model using human-generated reformulation data. CAPR innovatively incorporates a condition that mirrors user capability into the process of prompt reformulation. This design aligns seamlessly with our findings regarding the dependency of user reformulation on their capabilities. This unique approach ensures that CAPR is not excessively influenced by the inconsistent reformulation qualities prevalent in the training data. Instead, it enables the framework to adeptly learn diverse reformulation techniques from users with different levels of expertise. Furthermore, the condition representing user capability is adjustable, providing CAPR with the flexibility to function at a high level during inference. This allows CAPR to deliver high-quality prompt reformulations that surpass the average levels in its training dataset.

5.1 Model Architecture

In this section, we describe the overall architecture of CAPR. CAPR introduces the user capability into the reformulation process by decomposing a reformulation model into two components: a Conditional Reformulation Model (CRM) and Configurable Capability Features (CCF). The model architecture is illustrated in Figure 3. CCF is designed to represent various levels of user capacities in prompt writing, while CRM specializes in reformulating prompts in accordance with the specified capability. Mathematically, we denote CRM by ω and CCF by c. The reformulation process Ω is thus decomposed as follows:

$$\Omega(\hat{p}|p) = \sum_{c} \omega(\hat{p}|p, c) \cdot \mathrm{P}(c) \tag{2}$$

Here, p represents the original prompt, and $\hat{p}$ denotes the reformulated prompt. During training, by introducing user capability c as an additional condition, CRM (ω) can adapt to the various qualities of reformulation pairs and effectively learn different reformulation strategies. During inference, by configuring the distribution of CCF ($\mathrm{P}(c)$), CCF can enable CRM to simulate a high-capability user to reformulate prompts, transcending the limitations of the training data and generating reformulations with high expertise. Next, we describe the specific model designs.

5.1.1 Conditional Reformulation Model (CRM). CRM should effectively interpret the input user capability condition and reformulate the prompt accordingly. For this purpose, we implement it as a large language model due to its remarkable conditional generation abilities. As illustrated in Figure 3, the initial prompt and the user capability condition are transformed into textual formats, termed "meta prompts". These meta prompts are then concatenated and fed into a large language model to ensure that the task's nature and input details are clearly described, enabling the language model to accurately comprehend the conditions and produce conditional outputs.

5.1.2 Configurable Capability Features (CCF). CCF reflects the user's capability to effectively reformulate prompts. Given that such capacities are not explicitly recorded in interaction logs, CCF should be designed to be computable from reformulation pairs. Besides, it should also be easily tunable to guide the CRM toward simulating high-quality reformulations during inference. For these purposes, we employ a suite of scoring models to evaluate the quality of both the initial and reformulated prompts and use the output scores as CCF. This approach not only offers a direct assessment of the reformulation capability but also facilitates the straightforward extraction of user capacities from the interaction logs. The clarity and quantifiable nature of these metrics ensure they can be easily tuned during inference to lead CRM toward high-level reformulations. Specifically, CCF encompasses the following features:

- **Overall quality**: It measures the likelihood of user satisfaction with images generated from their prompts, serving as an indicator of overall prompt-reformulation ability. ImageReward model [51] is used to predict user satisfaction levels based on the generated results.
- **Prompt-image similarity**: It assesses the ability of a user's prompt to yield coherent images, a key aspect of generation quality. The CLIP model [38] is employed to assess the coherence between the prompt and the rendered image.
- **Aesthetic quality**: It indicates the visual appeal of the images produced from the prompts, a key aspect of generation quality. An aesthetic predictor [41] is used to evaluate the visual appeal of the generated images.
- **Prompt Length**: It reflects the user's skill in creating detailed prompts, a key aspect of prompt-writing capability. It is measured by the number of comma-separated phrases.

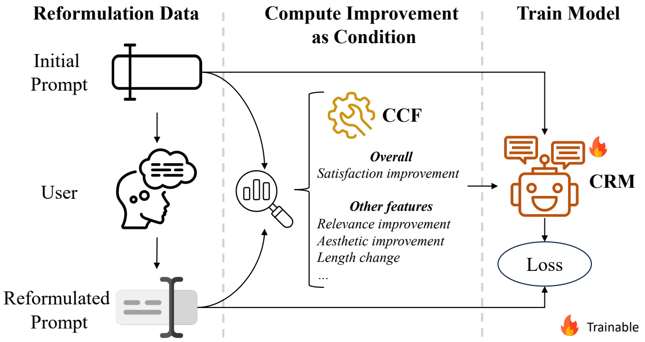


Figure 4: Training process of Capability-aware Prompt Reformulation (CAPR). Configurable Capability Features (CCF) is computed based on the training pairs, and Conditional Reformulation Model (CRM) is trained to predict the reformulated prompt given the initial prompt and CCF.

5.1.3 *Integrating CRM and CCF.*

To effectively integrate CRM with CCF, we construct a "meta prompt" using a structured template. This template is designed to contextualize the numeric features, making them more interpretable for the language model. The template used in our experiments is:

"A text-to-image generation system transforms text prompts into visual images. The effectiveness of this conversion depends on the prompt. The original prompt leads to images with prompt-image similarity of {}, aesthetic quality of {}, and overall quality of {}. To improve these metrics, new images are generated based on a revised prompt. After evaluating the new images for the initial prompt, the updated scores are: prompt-image similarity of {}, aesthetic quality of {}, and overall quality of {}. The revised prompt is structured into {} phrases, each separated by a comma. Considering the given information, the revised prompt should be:"

Next, we present the details of training the CAPR model.

5.2 Learning Conditional Generation

In this subsection, we detail the training process of the Conditional Reformulation Model (CRM). As illustrated in Figure 4, CRM training has two distinct stages, which are introduced below.

5.2.1 *Computing Configurable Capability Features (CCF).*

CRM's training requires the creation of data triplets: an initial prompt (p), its reformulation ($\hat{p}$), and a corresponding capability condition (c). While p and $\hat{p}$ are directly sourced from user inputs within the interaction logs, c is derived from each reformulation pair. As outlined in the previous subsection, we employ a suite of scoring models to assess the quality of both the initial and reformulated prompts, utilizing these evaluations as the basis for CCF.

The derived scores for overall quality, prompt-image similarity, and aesthetic quality are initially in floating-point formats with diverse ranges. To facilitate the language model's interpretation of these metrics, we convert them into a uniform integer scale. This is achieved by quantizing the scores into K integer values, ranging from 0 to $K-1$. The quantization is executed by evenly distributing the range of scores from the minimum to the maximum into K discrete intervals. In our experiments, we find CRM is robust to different K and empirically set K to 10.

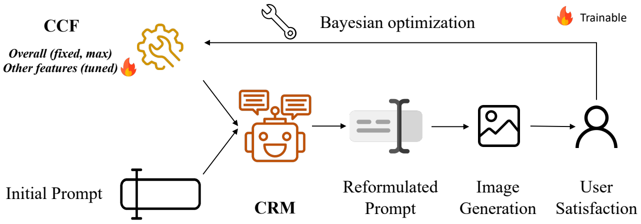


Figure 5: Configuration of Configurable Capability Features (CCF). The Conditional Reformulation Model (CRM) has been trained and is frozen. Within CCF, the overall quality metric is set to the highest, and other features of CCF are tuned to maximize the generation quality. The tuning process is accelerated with Bayesian optimization.

5.2.2 *Training Conditional Reformulation Model (CRM).*

The training of the CAPR model is guided by an autoregressive language modeling loss function. For each triplet of p, $\hat{p}$, and c, the model is trained to predict $\hat{p}$ given p and c. Mathematically, it is trained to minimize the following loss:

$$\mathcal{L} = -\log \omega(\hat{p}|p, c) = -\sum_{n} \log \omega(\hat{p}_n|\hat{p}_{1:n-1}, p, c) \tag{3}$$

where $\hat{p}_n$ is the n-th token of prompt $\hat{p}$. This process makes CAPR to learn different kinds of reformulation skills and accurately reformulate prompts according to the specified condition.

5.3 Configuring Capability Features

After completing the training for CRM, we focus on optimizing the Configurable Capability Features (CCF) to enhance reformulation performance. As introduced in Section 5.1.2, CCF for each reformulation pair incorporates two dimensions: quality assessments of the initial prompt and the reformulated prompt, referred to as c' and c'', respectively. These two components together form the composite CCF, denoted as:

$$c = (c', c'') \tag{4}$$

Next, we describe how to configure them, respectively.

5.3.1 *Efficient Prediction of Initial Prompt Quality.*

The assessment of the initial prompt's quality c' is executed efficiently within our framework. While the ideal method would involve generating and evaluating images based on the initial prompt, such a procedure is not practically feasible during inference because of the time-consuming generation process. To address this, we train a RoBERTa-Large model to predict the prompt quality based simply on the prompt. The trained model directly predicts the prompt's overall quality, its similarity to generated images, and aesthetic appeal during inference. While this deviates slightly from the precise quality scores, it significantly accelerates the evaluation process by saving the generation time. In practice, it reduces the time required from about 10 seconds to a mere 10 milliseconds.

5.3.2 *Optimizing Expected Reformulation Quality.*

The other aspect of CCF is the expected quality for reformulated results c''. During the training phase, this aspect of CCF serves as a guide for CRM by indicating the desired quality level of the generated prompt. During inference, this feature should be carefully set so that CRM is guided to perform optimally.

In an ideal scenario, setting this feature to its maximum guides the CRM to produce the most optimal prompt possible. However, this is not practical in practice due to the limits of the training data. Since few reformulated prompts in the training data can simultaneously achieve the max scores in terms of all CCF features (overall quality, similarity, and aesthetic quality), setting c'' to the maximum value pushes CRM to an extremely out-of-distribution scenario and potentially compromises its performance.

Therefore, we search for the best c'' that can maximize user satisfaction. This is depicted in Figure 5 and can be mathematically formulated as follows:

$$\begin{aligned} c''^{*} &= \arg\max_{c} \phi(c'') \\ \phi(c'') &= \mathop{\mathbb{E}}_{p,\hat{p},i} \left[\omega(\hat{p}|p,(c',c'')) \cdot \mathcal{G}(i|\hat{p}) \cdot f(p,i)\right] \end{aligned} \tag{5}$$

Here, $\mathcal{G}$ represents the generation system, and f measures user satisfaction, with $\phi(c'')$ indicating the average performance for a specified condition c''. While it would be ideal to tailor c'' for each individual prompt, it is impractical due to time and resource constraints. To simplify this process, we adopt several strategies:

- Search beforehand: By constructing a validation set, we determine the best configuration for c'' prior to the inference stage and thus eliminate the need for repetitive searches during individual prompt evaluations.
- Adaptive Reparameterization: To accommodate the diverse nature of prompts while maintaining pre-inference searching efficiency, we reparameterize c'' as:

  $$c'' = c' + \delta \tag{6}$$

  δ represents the expected improvement for CRM. We search for the optimal δ instead of c''. This allows tailoring c'' to individual prompt qualities without a one-size-fits-all c''.
- Search Space Reduction: Given the meaningful nature of each feature within CCF, we can heuristically narrow the search space. With our primary goal being overall user satisfaction, we set the expected overall quality factor to its maximum while searching for other features.
- Advanced Search Techniques: We employ Bayesian optimization for its efficiency and effectiveness. This method models $\phi(c'')$ as a Gaussian process, dynamically exploring new condition values based on previous results to refine the search progressively. Interested readers can refer to the skopt toolkit [22] for more details.

Through these optimization strategies, we can narrow down the search space from around 10^4 to 10^3 and use Bayesian optimization to conduct an effective search within 50 calls.

It is also important to note that the determined c'' is optimized for general user satisfaction. However, users with specific requirements, such as a focus on aesthetic quality, can use the provided c'' as a starting point and further customize the capability values to suit their needs. The flexibility and interpretability of CCF allow for easy user-driven adjustments in prompt reformulation behavior.

6 EXPERIMENTS

6.1 Experimental Setup

6.1.1 Training Data. We use a large-scale interaction log named DiffusionDB [47] for training. The dataset covers the real interactions from the official Stable Diffusion Discord channel for half a month. In total, it logs 1.8 million interactions from 10k users. When the interaction log was constructed, the text-to-image generation system used Stable Diffusion 1.4 model [39] for generation. We split the interactions into sessions, which is detailed in Section 4.2.1. We use the initial and the last prompts of a session to construct reformulation pairs. In total, we construct 30k reformulation pairs.

6.1.2 Evaluation Setup. We conduct evaluations using the HPSv2 benchmark dataset [49]. The dataset includes a wide variety of prompts categorized into Anime, ConceptArt, and Painting. Each category contains 800 prompts. For each prompt, we generate four images to ensure a robust assessment. We employ automated scoring models for assessments, including ImageReward [51] and HPSv2 [49]. They are trained to mimic human preferences and have been demonstrated to be reliable in evaluating text-to-image generation quality. We use two text-to-image generation models to evaluate the reformulation effectiveness, including Stable Diffusion 1.4 (SD1.4) [39] and Stable Diffusion XL base 1.0 (SDXL) [37]. SD1.4 is exactly the system used when the interaction log was constructed and therefore is a seen system to our model. SDXL is a recently released state-of-the-art generation system and thus is an unseen system.

6.1.3 Baselines. We compare CAPR against a comprehensive set of reformulation models, including:

- **GPT3.5 & 4** [7, 33]: They are used via APIs to reformulate prompts. We modify a popular prompt template from the web [4] to guide them for this task. The prompt contains task descriptions and well-crafted prompt examples.
- **PromptistSFT** [21]: This reformulation model is trained on synthesized reformulation pairs. The researchers first crawl well-performing prompts from online websites where users share prompts. Then they use ChatGPT to rephrase these prompts to poor prompts. Finally, they train a language model to predict the original prompts from the poor prompts.
- **PR-All** [16]: This is a traditional reformulation model. It utilizes a sequence-to-sequence transformer to predict the reformulated prompt. Compared to CAPR, it is trained on the same training data except that it does not introduce a condition mechanism.
- **PR-Weighted**: Compared to PR-All, it resolves the data quality problem by weighting each training pair based on its quality. It uses the quality improvement measured by ImageReward [51] to weight loss.
- **PR-Filter**: Compared to PR-All, it resolves the data quality problem by filtering out the low-quality pairs. The quality is measured as the improvement of ImageReward [51] score. We tune and select the best filtering threshold.

Table 2: Reformulation performance on the seen system (SD1.4). ImageReward [52] and HPSv2 [49] serve as evaluation models (f in Eq. (1)) and numbers are the average output scores. * and † separately indicate that performance is significantly better than SD1.4 and SD1.4+PR-Filter at $p < 0.01$ level measured by ttest. CAPR significantly outperforms baselines.

Method	ImageReward			HPSv2		
	Anime	ConceptArt	Painting	Anime	ConceptArt	Painting
SD1.4 [39]	0.038	0.185	0.190	27.42	26.86	26.86
+ GPT3.5 [7]	-0.037	0.030	0.126	27.36	26.77	26.87
+ GPT4 [33]	-0.143	-0.024	0.030	27.29	26.71	26.76
+ PromptistSFT [21]	-0.140	-0.083	0.010	27.19	26.60	26.77
+ PR-All [16]	0.094^{*}	0.180	0.233^{*}	27.51^{*}	26.91^{*}	26.95^{*}
+ PR-Weighted	0.083^{*}	0.164	0.227^{*}	27.48^{*}	26.87	26.97^{*}
+ PR-Filter	0.092^{*}	0.197	0.241^{*}	27.48^{*}	26.91^{*}	26.98^{*}
+ CAPR	**0.152***†	**0.213**	**0.311***†	**27.56***†	**26.95***†	**27.04***†

Table 3: Performance on an unseen and advanced system (SDXL). ImageReward [52] and HPSv2 [49] serve as evaluation models (f in Eq. (1)) and numbers are the average output scores. * and † separately indicate that performance is significantly better than SDXL and SDXL+PR-Filter at $p < 0.01$ level measured by ttest. CAPR effectively transfers to this unseen system.

Method	ImageReward			HPSv2		
	Anime	ConceptArt	Painting	Anime	ConceptArt	Painting
SDXL [37]	0.992	0.903	0.907	28.37	27.46	27.52
+ GPT3.5 [7]	0.883	0.762	0.832	28.26	27.32	27.50
+ GPT4 [33]	0.831	0.743	0.794	28.28	27.34	27.43
+ PromptistSFT [21]	0.785	0.638	0.688	28.12	27.19	27.34
+ PR-All [16]	1.014	0.926^{*}	0.948^{*}	28.43^{*}	27.51^{*}	27.61^{*}
+ PR-Weighted	1.008	0.919	0.941^{*}	28.44^{*}	27.52^{*}	27.62^{*}
+ PR-Filter	1.025^{*}	0.918	0.947^{*}	28.44^{*}	27.53^{*}	27.61^{*}
+ CAPR	**1.069***†	**0.949***†	**1.023***†	**28.50***†	**27.56***†	**27.68***†

Note that except for GPT3.5 & 4, the remaining baselines are all trained as a simple sequence-to-sequence translation model, as in previous reformulation models in web search scenarios [16]. They do not employ a condition in the model designs and simply view reformulation as a source-target translation task.

6.1.4 Implementation Details.

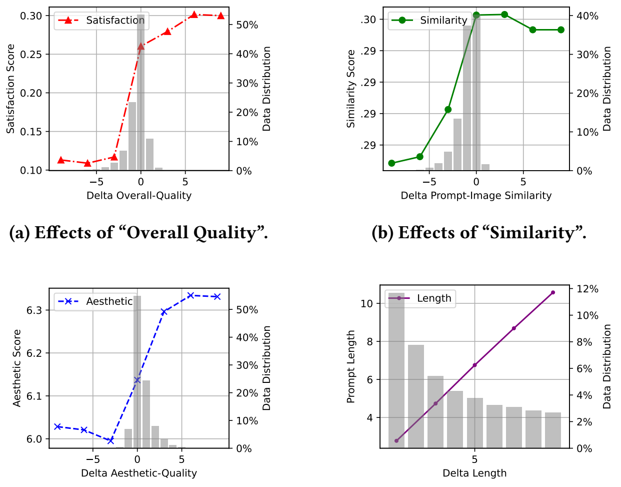
CRM is initialized with TinyLlama [53], a 1.1B model trained on 2.5T tokens. Training lasts for 2 epochs with AdamW [28] optimizer and a learning rate of 4×10^{-5}. When tuning CCF, we construct a validation set of 100 prompts and set the inference steps of SD1.4 to 20 steps for acceleration. We use the gp_minimize function from skopt package [22] to efficiently search the optimal CCF values within 50 generation calls. For more details, please refer to our open-sourced code.

6.2 Experimental Results

In this section, we present the experimental results. The experiments are conducted on two text-to-image generation systems, namely SD1.4 and SDXL. SD1.4 is the model that corresponds to the training interaction log and thus has been seen by CAPR and baselines, while SDXL has not been seen by the reformulation models and thus can evaluate the model robustness.

6.2.1 Performance on the Seen System. The performance on SD1.4 is shown in Table 2. Results demonstrate that CAPR substantially improves the generation quality and significantly outperforms all baselines. Specifically, we have the following observations:

- Generic language models like GPT3.5 and GPT4 cannot improve the generation quality. After manually examining the results, we find that although both models try to mimic the formulation of well-crafted prompts, they tend to hallucinate and change the prompt meanings. Besides, although the output of GPT4 is more organized than GPT3.5 in terms of prompt structure, GPT4 adds a lot of modifier words like image style that misalign with user intention. This adversarially leads to worse results than GPT3.5.
- According to the performance of PromptistSFT, training on synthetically generated refinement pairs results in limited effectiveness. In its training data, the target labels are well-crafted prompts crawled from the web, while the user inputs are simulated by rephrasing these prompts to everyday language with ChatGPT. Nevertheless, the distribution of such rephrased text is different from that of real users' input, resulting in an ineffective model.
- Training on users' reformulation data helps models learn to reformulate. PR-All, PR-Weighted, and PR-Filter all improve the generation quality. We also observe that PR-Weighted and PR-Filter perform similarly to PR-All. Although both models preprocess the training data to focus on training data with high quality, this preprocessing also results in limited training data size, which adversarially affects model training.



(a) Effects of "Overall Quality".

(b) Effects of "Similarity".

(c) Effects of "Aesthetic Quality".

(d) Effects of "Prompt Length".

Figure 6: Effects of CCF conditions for CRM. X-axis is the input expected performance improvement, as formulated in Eq. (6). Lines show the evaluation output quality, and bars show the training data distribution. Results suggest that CRM can be effectively controled by the input condition.

- CAPR significantly outperforms the baselines. Compared with generic language models and PromptistSFT, CAPR is trained on real users' reformulation data and effectively learns useful reformulation strategies. Compared with PR-All/Weighted/Filter, CAPR adopts a conditional reformulation framework that can better address the quality issue of user reformulation data.

6.2.2 Performance on the Unseen System. Table 3 shows the performance on SDXL [37], an advanced model not used during training. Results suggest that CAPR effectively transfers to this new model and significantly improves the generation performance. We have the following observations:

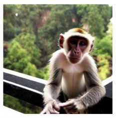
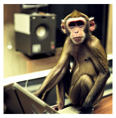
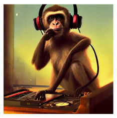
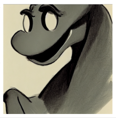
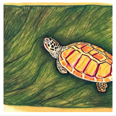
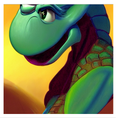
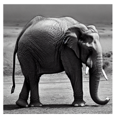
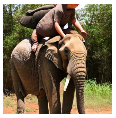
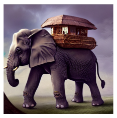
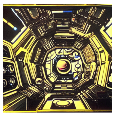
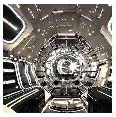
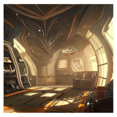
- SDXL can evaluate the robustness of our reformulation models because it is substantially more advanced than SD1.4. According to Table 2 and Table 3, SDXL substantially improves the generation performance of SD1.4. Since the interaction log used for training is for SD1.4, SDXL can evaluate how reformulation models generalize to more advanced models.
- CAPR still significantly improves the generation performance of SDXL and outperforms all baselines. This demonstrates that the benefits of CAPR are parallel to the advance of text-to-image generation system.

6.3 Model Analysis

The key of CAPR is to guide the reformulation behaviour using configurable feature conditions (CCF). In this section, we deeply analyze how different conditions control the CAPR performance.

The methodology is to tune each factor separately and observe the generation quality. We use the "Painting" dataset for analysis. We tune the expected reformulation quality, as formulated in Eq. (6), i.e., requiring CAPR to improve a factor by a specified delta value (δ in Eq. (6)). The optimal delta values learned from the validation data (described in Section 5.3.2) are 9 for overall quality, 9 for aesthetic quality, 0 for prompt-image similarity, and 5 for prompt length. To analyze the effects of each factor, we freeze other factors to eliminate their influence: keeping the delta values for prompt length to 5 and other factors to 0. In the following, we discuss how CAPR follows each condition factor.

Table 4: Prompt reformulation examples. Column one details user prompts and the reformulated prompts by PR-Filter and CAPR. The next three columns show images generated by SD1.4 from user inputs or reformulated prompts. CAPR substantially enhances the image quality.

Input Prompt	User Input	PR-Filter	CAPR
User Input: A monkey is pictured acting as a DJ. **PR-Filter**: A monkey DJ. **CAPR**: A monkey, wearing headphones, A monky is pecturing as a dj., digital art, artstation, by greg rutkowski			
User Input: Milt Kahl's sketch of Cecil Turtle. **PR-Filter**: Milt Kahl sketch of a turtle. **CAPR**: Milt Kahl's sketch cecil turtle. detailed, high quality, digital painting, fantasy, artwork, in the style of Cecilia Turtles			
User Input: An elephant carrying a house on its back. **PR-Filter**: An elephantine elephantin carrying a human on its shoulders. **CAPR**: An elephant carrying a House on its Back. Fantasy, digital painting, HD, 4k, detailed, artwork			
User Input: The interior of a spaceship orbiting alpha centauri. **PR-Filter**: The interior of an alien spaceship. **CAPR**: The interior of spaceship of a fantasy setting, highly detailed, digital painting, artstation, concept art, illustration			

As shown in Figure 6, CAPR can follow the specified condition well and even extrapolate beyond the training data distribution. Specifically, when we gradually increase the delta values, the corresponding metric scores improve. This demonstrates that CCF can control the behaviour of CAPR. In Figure 6, we also depict the distribution of delta scores for each CCF factor in the training data (i.e., the grey bars in the figures). We can see that most user-reformulated prompts have near-zero delta scores in terms of different image quality measurements. This low-quality training data is primarily caused by the fact that users' capacity is relatively stable during one session, as discussed in Section 4. Nevertheless, CAPR can extrapolate beyond the limits of the training. For example, although few training pairs improve the overall quality more than 2, as shown in Figure 6a, CAPR can still improve user satisfaction when the condition is increased from 2 to 6.

6.4 Case Studies

Table 4 includes four reformulation examples. We can see that PR-Filter only slightly rephrases user inputs and even hallucinates in the third case. This stems from the training data where human-generated reformulation pairs are alike, as discussed in Section 4.

Instead, CAPR reformulates prompts to keyword-enriched ones with artist names and stylistic elements, which have been demonstrated to be favored by text-to-image generation systems [35, 48]. For instance, in the first example about a monkey DJ, CAPR adds the "headphones" detail and the artist name "greg rutkowski". These cases demonstrate that the conditional framework is the key to learning prompt reformulation from interaction logs.

7 CONCLUSION

Text-to-image generation systems have increasingly become a milestone in digital art creation. Yet, their effectiveness is closely tied to the quality of the prompts provided by users, a task that often presents significant challenges to the average user. In this paper, we leverage user interaction logs as a valuable resource for training an automatic prompt reformulation model. Our investigation reveals a distinctive aspect of prompt reformulation in text-to-image systems: it relies heavily on the user's intrinsic ability to craft effective prompts rather than on the system's feedback. To address this unique challenge, we introduce the Capability-aware Prompt Reformulation (CAPR) framework, a pioneering solution for training on interaction logs. CAPR can adapt to various user capabilities and simulate high-quality reformulation during inference. Extensive experiments demonstrate the effectiveness of CAPR, highlighting its significant improvements over existing baselines and transferability to unseen systems.

ACKNOWLEDGMENTS

This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301).

REFERENCES

[1] [n.d.]. PromptHero - Search prompts for Stable Diffusion, ChatGPT & Midjourney. https://prompthero.com/ Accessed: 2024-01-20.

[2] [n.d.]. Stable Diffusion - Prompts examples. https://stablediffusion-fr.webpkgcache.com/doc/-/s/stablediffusion.fr/prompts Accessed: 2024-01-20.

[3] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. 2004. Query recommendation using query logs in search engines. In *International conference on extending database technology*. Springer, 588–596.

[4] bluelovers. 2023. ChatGPT Stable Diffusion Prompts Generator. https://gist.github.com/bluelovers/92dac6fe7dcbafd7b5ae0557e638e6ef#file-chatgpt-stable-diffusion-prompts-generator-txt. Accessed: 2023-7-20.

[5] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. 2011. Query reformulation mining: models, patterns, and applications. *Information retrieval* 14 (2011), 257–289.

[6] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. 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*. 1–14.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Fei Cai, Maarten De Rijke, et al. 2016. A survey of query auto completion in information retrieval. *Foundations and Trends® in Information Retrieval* 10, 4 (2016), 273–363.

[9] Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)* 44, 1 (2012), 1–50.

[10] Michael Chau, Xiao Fang, and Olivia R Liu Sheng. 2005. Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology* 56, 13 (2005), 1363–1376.

[11] Jia Chen, Jiaxin Mao, Yiqun Liu, Ziyi Ye, Weizhi Ma, Chao Wang, Min Zhang, and Shaoping Ma. 2021. A Hybrid Framework for Session Context Modeling. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–35.

[12] Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. Towards a better understanding of query reformulation behavior in web search. In *Proceedings of the web conference 2021*. 743–755.

[13] Jing Chen, Dan Wang, Iris Xie, and Quan Lu. 2018. Image annotation tactics: transitions, strategies and efficiency. *Information Processing & Management* 54, 6 (2018), 985–1001.

[14] Jerry Zikun Chen, Shi Yu, and Haoran Wang. 2020. Exploring Fluent Query Reformulations with Text-to-Text Transformers and Reinforcement Learning. *arXiv preprint arXiv:2012.10033* (2020).

[15] Niklas Deckers, Julia Peters, and Martin Potthast. 2023. Manipulating Embeddings of Stable Diffusion Prompts. *arXiv preprint arXiv:2308.12059* (2023).

[16] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1747–1756.

[17] Efthimis N Efthimiadis. 2000. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science* 51, 11 (2000), 989–1003.

[18] Archan Ghosh, Debgandhar Ghosh, Madhurima Maji, Suchinta Chanda, and Kalporup Goswami. 2023. MTTN: Multi-Pair Text to Text Narratives for Prompt Generation. *arXiv preprint arXiv:2301.10172* (2023).

[19] Dongyi Guan, Sicong Zhang, and Hui Yang. 2013. Utilizing query change for session search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 453–462.

[20] Kishaloy Halder, Heng-Tze Cheng, Ellie Ka In Chio, Georgios Roumpos, Tao Wu, and Ritesh Agarwal. 2020. Modeling Information Need of Users in Search Sessions. *arXiv preprint arXiv:2001.00861* (2020).

[21] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611* (2022).

[22] Tim Head, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. 2021. *scikit-optimize/scikit-optimize*.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[24] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 445–454.

[25] Minsoo Kang, Doyup Lee, Jiseob Kim, Saehoon Kim, and Bohyung Han. 2023. Variational Distribution Learning for Unsupervised Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23380–23389.

[26] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up gans for text-to-image synthesis. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10124–10134.

[27] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational diffusion models. *Advances in neural information processing systems* 34 (2021), 21696–21707.

[28] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[29] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. 2022. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 18187–18196.

[30] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–23.

[31] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. 2023. Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models To Learn Any Unseen Style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 14267–14276.

[32] Midjourney. 2023. Midjourney: An Independent Research Lab Exploring New Mediums of Thought. https://www.midjourney.com/. [Online; accessed 21-January-2024].

[33] OpenAI. 2023. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2023-11-13.

[34] OpenAI. 2023. Improving Image Generation with Better Captions. https://cdn.openai.com/papers/dall-e-3.pdf. Accessed: 2023-11-13.

[35] Jonas Oppenlaender. 2022. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988* 2 (2022).

[36] Guy Parsons. 2022. The DALL·E 2 Prompt Book. https://dallery.gallery/the-dalle-2-prompt-book.

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PMLR, 8748–8763.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10684–10695.

[40] Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American society for information science* 41, 4 (1990), 288–297.

[41] Christoph Schuhmann. 2022. Improved Aesthetic Predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[43] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. 2023. Prompt Stealing Attacks Against Text-to-Image Generation Models. *arXiv preprint arXiv:2302.09923* (2023).

[44] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. In *Acm sigir forum*, Vol. 33. ACM New York, NY, USA, 6–12.

[45] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *proceedings of the 24th ACM international on conference on information and knowledge management.* 553–562.

[46] Amanda Spink, Bernard J Jansen, and H Cenk Ozmultu. 2000. Use of query reformulation and relevance feedback by Excite users. *Internet research* 10, 4 (2000), 317–328.

[47] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.* 893–911.

[48] Sam Witteveen and Martin Andrews. 2022. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462* (2022).

[49] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341* (2023).

[50] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. 2023. A prompt log analysis of text-to-image generation systems. In *Proceedings of the ACM Web Conference 2023.* 3892–3902.

[51] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977* (2023).

[52] Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2020. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848* (2020).

[53] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An Open-Source Small Language Model. arXiv:2401.02385 [cs.CL]

[54] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. 2023. Shifted diffusion for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10157–10166.