A Prompt Log Analysis of Text-to-Image Generation Systems

Yutong Xie*
yutxie@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Zhaoying Pan*
panzy@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Jinge Ma*
jingema@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Luo Jie
rogerluo@nianticlabs.com
Niantic, Inc.
San Francisco, California, USA

Qiaozhu Mei
qmei@umich.com
University of Michigan
Ann Arbor, Michigan, USA

ABSTRACT

Recent developments in large language models (LLM) and generative AI have unleashed the astonishing capabilities of text-to-image generation systems to synthesize high-quality images that are faithful to a given reference text, known as a "prompt". These systems have immediately received lots of attention from researchers, creators, and common users. Despite the plenty of efforts to improve the generative models, there is limited work on understanding the information needs of the users of these systems at scale. We conduct the first comprehensive analysis of large-scale prompt logs collected from multiple text-to-image generation systems. Our work is analogous to analyzing the *query logs* of Web search engines, a line of work that has made critical contributions to the glory of the Web search industry and research. Compared with Web search queries, text-to-image prompts are significantly longer, often organized into special structures that consist of the *subject*, *form*, and *intent* of the generation tasks and present unique categories of information needs. Users make more edits within creation sessions, which present remarkable *exploratory* patterns. There is also a considerable gap between the user-input prompts and the captions of the images included in the open training data of the generative models. Our findings provide concrete implications on how to improve text-to-image generation systems for creation purposes.

KEYWORDS

Text-to-Image Generation, AI-Generated Content (AIGC), AI for Creativity, Prompt Analysis, Query Log Analysis.

ACM Reference Format:

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 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3587430

*These authors contributed equally to this research.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3587430

1 INTRODUCTION

Recent developments in large language models (LLM) (e.g., GPT-3 [4], PaLM [6], LLaMA [38], and GPT-4 [24]) and generative AI (especially the diffusion models [13, 37]) have enabled the astonishing image synthesis capabilities of text-to-image generation systems, such as DALL·E [29, 30], Midjourney [20], latent diffusion models (LDMs) [32], Imagen [33], and Stable Diffusion [32]. As these systems are able to produce images of high quality that are faithful to a given reference text (known as a "prompt"), they have immediately become a new source of creativity [25] and attracted a great number of creators, researchers, and common users. As a major prototype of generative AI, many believe that these systems are introducing fundamental changes to the creative work of humans [9].

Despite plenty of efforts on improving the performance of the underneath generative models, there is limited work on analyzing the information needs of the real users of these text-to-image systems, regardless of the cruciality to understand the objectives and workflows of the creators and identify the gaps in how the current systems are capable of facilitating the creators' needs.

In this paper, we take the initiative to investigate the information needs of text-to-image generation by conducting a comprehensive analysis of millions of user-input prompts in multiple popular systems, including Midjourney, Stable Diffusion, and LDMs. Our analysis is analogous to *query log analysis* of search engines, a line of work that has inspired many developments of modern information retrieval (IR) research and industry [3, 12, 14, 36, 44]. In this analogy, a text-to-image generation system is compared to a search engine, the pretrained large language model is compared to the search index, a user-input prompt can be compared to a search query that describes the user's information need, while a text-to-image generation model can be compared to the search or ranking algorithm that generates (rather than retrieves) one or multiple pieces of content (images) to fulfill the user's need (Table 1).

Through a large-scale analysis of the prompt logs, we aim to answer the following questions: (1) *How do users describe their information needs in the prompts*? (2) *How do the information needs in text-to-image generation compare with those in Web search*? (3) *How are users' information needs satisfied*? (4) *How are users' information needs covered by the image captions in open datasets*?

The results of our analysis suggest that (1) text-to-image prompts are usually structured with terms that describe the *subject*, the *form*, and the *intent* of the image to be created (Sec. 4); (2) text-to-image prompts are sufficiently different from Web search queries. Besides

Table 1: The analogy between text-to-image generation and Web or vertical search.

Text-to-image generation	Web and vertical search
Images	Webpages/Documents
Generation	Retrieval
Text-to-image generation system	Search engine
Prompt	Query
Pretrained language model	Search index
Image generation models	Ranking algorithms
Prompt log analysis	Query log analysis
...	...

significantly lengthier prompts and sessions, there is especially a prevalence of *exploratory* prompts (Sec. 4.2); (3) image generation quality (measured by user rating) is correlated with the length of the prompt as well as the usage of terms (Sec. 4.3); and (4) there is a considerable gap between the user-input prompts and the image captions in open datasets (Sec. 4.4). More details of our analysis are listed in the Appendix, and the code and the complete results are accessible via our GitHub repository[1]. Based on these findings, we conclude several challenges and actionable opportunities of text-to-image generation systems (Sec. 5). We anticipate our study would help the text-to-image generation community to better understand and facilitate creativity on the Web.

2 RELATED WORK

2.1 Text-to-Image Generation

Text-to-Image generation is a multi-modal task that aims to translate text descriptions (known as "prompts") into faithful images of high quality. Recent text-to-Image generation models could be categorized into two main streams: (1) models based on variational autoencoders (VAEs) [16] or generative adversarial networks (GANs) [11], and (2) models built upon denoising diffusion probabilistic models (DDPMs, or diffusion models) [13].

The earliest application of deep neural networks in text-to-image generation could be dated back to 2015, where Mansimov et al. [18] proposed to generate images from texts using a recurrent VAE with the attention mechanism. In the next few years, Reed et al. [31] and Cho et al. [5] started to use GANs as generative models from texts to images. These models have made it possible to generate images from texts, however, most generated images are blurry and consist of simple structures. Later in 2021, OpenAI released DALL·E, combining the powerful GPT-3 language model [4] as the text encoder and a VAE as the image generator [30]. DALL·E is able to generate more complex and realistic images, establishing a new standard of text-to-image generation.

Since late 2021, with the advances of DDPMs (diffusion models), several compelling text-to-image generation systems have been developed and released to the public, demonstrating astounding capabilities in faithful image synthesis and bringing text-to-image generation into a new era. They include Disco Diffusion[2], GLIDE [21], Midjourney [20], DALL·E 2 [29], latent diffusion models (LDMs) [32], Imagen [33] and Stable Diffusion [32]. These systems immediately become a trend in the art creation community, attracting both artists and common users to create with such systems [25].

2.2 Text-to-Image Prompt Analysis

Despite plenty of efforts on improving the performance of the underneath generative models, there is limited work on analyzing the user-input prompts and understanding the information needs of the real users of text-to-image systems.

Liu and Chilton [17] explored what prompt keywords and model hyperparameters can lead to better generation performance from the human-computer interaction (HCI) perspective. In particular, 51 keywords related to the subject and 51 related to the style have been tested through synthetic experiments. Oppenlaender [26] further conducted an autoethnographic study on the modifiers in the prompts. As a result, six types of prompt modifiers have been identified, including subject terms, style modifiers, image prompts, quality boosters, repetition, and magic terms. In addition, Pavlichenko and Ustalov [27] presented a human-in-the-loop approach and extracted some most effective combinations of prompt keywords.

These studies have provided valuable insights into certain aspects of text-to-image prompts. However, these researches are mostly based on small numbers of independent prompts and/or computational experiments. These lab experiments usually do not consider prompts in real usage sessions and can hardly reflect the "whole picture" of the information needs of the real users. Our study provides the first large-scale quantitative analysis based on the user inputs collected from real systems. Besides, we also compare the characteristics of prompts with those of Web search queries as well as image captains in open text-image datasets, revealing considerable differences and practical implications.

2.3 Query Log Analysis

Query log analysis of Web search engines is a classical line of work that has inspired many developments in modern information retrieval (IR) research and industry. Such an analysis usually includes examinations into terms, queries, sessions, and users [3, 14, 36]. Aside from the general research on Web search engines, query log analysis has also been conducted on vertical search engines like medical search engines (*e.g.*, PubMed and electronic health records (EHR) search engine), where the analysis results are further compared with Web search patterns [12, 44].

In this paper, we make an analogy between query log analysis and prompt analysis. In this analogy, a user-input prompt can be compared to a *search query* that describes the user's information need, while a text-to-image generation system could be compared to a *search engine* that generates (rather than retrieves) one or more pieces of content (in our case, the image(s)) to fulfill the user's need.

3 PROMPT LOG DATASETS

We consider three large and open prompt log datasets, including the Midjourney Discord dataset [39], the DiffusionDB [40], and the Simulacra Aesthetic Captions (SAC) [28]. These datasets involve three popular text-to-image generation systems – Midjourney [20], Stable Diffusion [32], and latent diffusion models (LDMs) [32].

[1]GitHub repository: https://github.com/zhaoyingpan/prompt_log_analysis.
[2]Disco Diffusion: https://github.com/alembics/disco-diffusion, retrieved on 3/14/2023.

Table 2: Statistics of datasets. The values (except for the raw number of records) are calculated after data processing.

Dataset	Midjourney	DiffusionDB	SAC
Raw #Records	250K	14M	238K
#Prompts	145,074	2,208,019	34,190
#Unique prompts	122,905	1,817,721	34,190
#Unique terms	97,052	182,386	22,898
#Users	1,665	10,380	N/A
Median #prompts/user	12	62	N/A
Max #prompts/user	2,493	19,556	N/A

The Midjourney Discord dataset. The Midjourney dataset [39] is obtained by crawling message records from the Midjourney Discord community over a period of four weeks (June 20 – July 17, 2022). This dataset contains approximately 250K records, with user-input prompts, URLs of generated images, usernames, user IDs, message timestamps, and other Discord message metadata.

DiffusionDB. DiffusionDB [40] is a large-scale dataset with 14M images generated with Stable Diffusion [32]. For each image, this dataset also provides the corresponding prompt, user ID, timestamp, and other meta information.

Simulacra Aesthetic Captions (SAC). The SAC dataset [28] contains 238K images generated from over 40K user-submitted prompts with LDMs [32]. SAC annotates images with aesthetic ratings in the range of $[1, 10]$ collected from surveys. The prompts in SAC are also relatively clean. However, SAC does not include information about user IDs or timestamps.

Table 2 lists the basic statistics of the datasets. In the raw data, one input prompt can correspond to multiple generated images and create multiple data entries for the same input. We remove these duplicates while reserving repeated inputs from users. More details about the data and data processing are described in Appendix A.

4 PROMPT LOG ANALYSIS

We analyze the prompts in the datasets and aim to answer the four questions mentioned in Section 1.

4.1 How do Users Describe Information Needs?

We first investigate how users describe their information needs by exploring the structures of prompts. We start with analyzing the usage of terms (tokens or words) in prompts. We conduct a first-order analysis that focuses on term frequency, followed by a second-order analysis that focuses on co-occurring term pairs. The significance of a term pair is measured with the χ^2 metric [1, 36]:

$$\chi^2(a,b) = \frac{[E(ab) - O(ab)]^2}{E(ab)} + \frac{[E(\bar{a}b) - O(\bar{a}b)]^2}{E(\bar{a}b)} + \frac{[E(a\bar{b}) - O(a\bar{b})]^2}{E(a\bar{b})} + \frac{[E(\bar{a}\bar{b}) - O(\bar{a}\bar{b})]^2}{E(\bar{a}\bar{b})}, \quad (1)$$

where a, b are two terms, $O(ab)$ is the number of prompts they co-occur in, $E(ab)$ is the expected co-occurrences under the independence assumption, and $\bar{a}$, $\bar{b}$ stand for the absence of a, b.

In Table 3, we list the most frequent terms, measured by the number of text-to-image prompts they appear in. The most significant term pairs are listed in Table 4. Based on the first- and second-order analysis results, we present the following findings.

Table 3: Most frequent terms used in prompts.

	Midjourney		DiffusionDB		SAC	
	Term	**Freq.**	**Term**	**Freq.**	**Term**	**Freq.**
1	,	91,993	,	1,689,552	,	17,265
2	>	71,353	of	1,084,836	of	15,400
3	<	71,348	a	1,043,542	a	14,693
4	of	56,470	by	943,503	by	12,442
5	a	47,987	and	720,802	the	9,319
6	in	42,685	in	669,648	and	7,614
7	--ar	40,014	detailed	653,587	in	7,186
8	the	38,155	art	598,493	on	6,723
9	and	33,330	the	572,569	artstation	5,935
10	by	28,074	artstation	484,898	.	5,686
11	detailed	25,134	on	475,406	portrait	5,652
12	with	24,112	painting	426,930	art	5,598
13	style	23,461	portrait	412,547	with	4,444
14	on	20,282	with	402,008	painting	4,347
15	render	20,061	highly	334,410	illustration	3,359
16	cinematic	19,782	k	320,290	-	3,354
17	16:9	18,616	lighting	310,598	oil	3,234
18	realistic	18,012	digital	310,336	concept	3,214
19	-	17,677	-	287,732	digital	2,997
20	octane	16,925	intricate	276,246	beautiful	2,695

Table 4: Most significant term pairs used in the same prompt.

	Midjourney		DiffusionDB		SAC	
	Pair	χ^2	**Pair**	χ^2	**Pair**	χ^2
1	(norman, rockwell)	0.285	(donald, trump)	0.345	(matsunuma, shingo)	1.000
2	(fenghua, zhong)	0.250	(emma, watson)	0.321	(lisa, mona)	1.000
3	(ngai, victo)	0.240	(biden, joe)	0.283	(elon, musk)	1.000
4	(makoto, shinkai)	0.237	(shinkawa, yoji)	0.266	(ariel, perez)	1.000
5	(ray, trace)	0.125	(blade, runner)	0.255	(angeles, los)	1.000
6	(fiction, science)	0.123	(katsuhiro, otomo)	0.238	(bradley, noah)	1.000
7	(anderson, wes)	0.106	(contest, winner)	0.237	(hayao, miyazaki)	1.000
8	(11:17, circa)	0.074	(takato, yamamoto)	0.236	(finnian, macmanus)	1.000
9	(jia, ruan)	0.071	(“, ”)	0.216	(bartlett, bo)	1.000
10	(cushart, krenz)	0.070	(mead, syd)	0.130	(hasui, kawase)	0.500
11	(shinkawa, yoji)	0.062	(akihiko, yoshida)	0.123	(daniela, uhlig)	0.332
12	(albert, bierstadt)	0.060	(elvgren, gil)	0.114	(edlin, tyler)	0.318
13	(katsuhiro, otomo)	0.057	(new, york)	0.114	(jurgens, mandy)	0.286
14	([,])	0.053	(gi, jung)	0.106	(bacon, francis)	0.286
15	(annie, leibovitz)	0.052	(dore, gustave)	0.103	(araki, hirohiko)	0.258
16	(adams, ansel)	0.045	(star, wars)	0.092	(radke, scott)	0.257
17	(mignola, mike)	0.043	(fiction, science)	0.087	(ca', n't)	0.252
18	(1800s, tintype)	0.036	(league, legends)	0.082	(card, tarot)	0.201
19	(dore, gustave)	0.036	(rule, thirds)	0.074	(claude, monet)	0.190
20	(adams, tintype)	0.029	(ngai, victo)	0.061	(gogh, van)	0.180

4.1.1 ***Words in prompts describe subjects, forms, and intents.***

In Art, a piece of work is typically described with three basic components: *subject*, *form*, and *content*. In general, the *subject* defines "what" (the topic or focus); the *form* confines "how" (the development, composition, or substantiation); and the *content* articulates "why" (the intention or meaning) [22]. We are able to relate terms in a text-to-image prompt to these three basic components. Note that the *subject*, *form*, and *content* of a work of art is often intertwined with each other. For example, a term describing the *subject* might also be related to the *form* or *content* and vice versa.

Subject. A prompt often contains terms describing its topic or focus, referred to as the *subject*, which can be a person, an object, or a theme [26, 27]. Among the 50 most frequent terms of all three datasets (parts of them listed in Table 3), we discover 9 terms related to the *subject*: “portrait”, “lighting”, “light”, “face”, “background”, “character”, “man”, “head”, and “space”. More examples can be found in Table 4, such as (“donald”, “trump”), (“emma”, “watson”), (“biden”, “joe”), (“elon”, “musk”), (“mona”, “lisa”), (“new”, “york”), (“los”, “angeles”), (“star”, “wars”), (“league”, “legends”), and (“tarot”, “card”).

Form. The *form* confines the way in which an artwork is organized, referring to the use of the *principles of organization* to arrange the elements of art. These elements may include *line*, *texture*, *color*, *shape*, and *value*; while the principles of organization consider *harmony*, *variety*, *balance*, *proportion*, *dominance*, *movement*, and *economy*, *etc.* [22]. Comparably, the *form* of a prompt is usually described as *constraints* to image generation [26, 27]. Among the top 50 terms of all datasets (parts of them listed in Table 3), we find 25 terms that are *form*-related: “detailed”/“detail”', “art”, “painting”, “style”, “render”, “illustration”, “cinematic”, “k” (*e.g.*, “4K” or “8K”), “16:9”/“9:16”, “oil” (*e.g.*, “oil painting”), “realistic”, “concept” (*e.g.*, “concept art”), “digital”, “intricate”, “black”, “dark”, “unreal”, “white”, “sharp”, “fantasy”, “photo”, “smooth”, and “canvas”.

In addition to these terms, we also notice names of art community Websites (*e.g.*, ArtStation[3], Artgerm[4], and CGSociety[5]), rendering engines (*e.g.*, Unreal Engine[6] and OctaneRender[7]), and artists (*e.g.*, wlop, Norman Rockwell, Fenghua Zhong, Victo Ngai, Shingo Matsunuma, Claude Monet, and Van Gogh, *etc.*) that appear frequently in the prompts (Tables 3-4). These terms are often used to constrain the style of images, so can be interpreted as *form*-related.

Intent. The *content* (as defined in the Art literature) of a prompt tells the intention or purpose of the user and is often described as the emotional or intellectual message that the user wants to express. Among the three components of art, the *content* is the most abstract and can be difficult to identify [22]. To avoid ambiguity (“content” has specific meanings in the Web and the AI literature), we name this component of a prompt the “*intent*” instead. In the top 50 terms of all datasets (parts of them listed in Table 3), we find only three terms that might be related to the *intent*: “beautiful”, “trending”, and “featured”. If we go down the list, we are able to identify more: “epic”, “moody”, “fantasy”, “dramatic”, “masterpiece”, *etc.*

Other terms. Aside from the terms that describe the *subject*, *form*, and *intent*, other types of frequently used terms include punctuations (*e.g.*, “,” and “.”), model-specific syntactic characters (*e.g.*, “<”, “>”, “--ar”, and “::” that specify model parameters in the Midjourney dataset), and stop words (*e.g.*, “of”, “the”, “in”, “a”, “and”, and “by”).

Overall, we find that many of the prompts consist of one or more blocks of terms in at least one of these three categories. The frequent appearance of *form*-related terms is particularly interesting, which adds constraints to the creation process. Future developments of text-to-image generation should consider how to optimize for the users' intents under these constraints.

[3]ArtStation: https://www.artstation.com/, retrieved on 3/14/2023.
[4]Artgerm: https://artgerm.com/, retrieved on 3/14/2023.
[5]CGSociety: https://cgsociety.org/, retrieved on 3/14/2023.
[6]Unreal Engine: https://www.unrealengine.com/, retrieved on 3/14/2023.
[7]OctaneRender: https://home.otoy.com/render/octane-render/, retrieved on 3/14/2023.

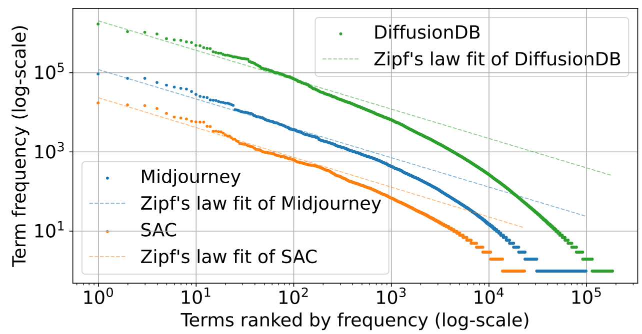

Figure 1: Term frequencies in a log-log scale. The distributions deviate from Zipf's law with exponential tails.

4.1.2 ***Prompts indicate potential applications.*** In the second-order analysis, we also discover interesting combinations of terms that might indicate potential applications of text-to-image generation in various areas:

- *Image processing*: (“film”, “grain”), (“blur”, “blurry”), (“iso”, “nikon”), (“hdr”, “professional”), (“flambient”, “professional”), (“post”, “processing”), (“color”, “scheme”), *etc.*
- *Image rendering*: (“ray”, “trace/tracing”), (“fluid”, “redshift”), (“unreal”, “engine”), (“3d”, “shading”), (“3d”, rendering”), (“global”, “illumination”), (“octane”, “render”), *etc.*
- *Graphic design*: (“movie”, “poster”), (“graphic”, “design”), (“key”, “visual”), (“cel”, “shaded”), (“comic”, “book”), (“anime”, “visual”), (“ghibli”, “studio”), (“disney”, “pixar”), *etc.*
- *Industrial design*: (“circuit”, “boards”), (“sports”, “car”), *etc.*
- *Fashion design*: (“fashion”, “model”), (“curly”, “hair”), *etc.*

These pairs are often related to the *forms* or/and *intents* of the creation, indicating considerable opportunities to develop customized applications for different forms and intents of creative activities.

4.2 How do Text-to-Image Prompts Compare with Web Search Queries?

A text-to-image prompt is analogous to a *query* submitted to a Web search engine (image generation model) that retrieves (generates) documents (images) that satisfy the information need (Table 1). It is intriguing to compare text-to-image prompts with Web search queries to further understand their similarities and differences.

4.2.1 ***Term frequencies do not follow the power law.*** While a power law distribution (or a Zipf's distribution when the rank of terms is the independent variable) of term frequency is commonly observed in large-scale corpora and Web search queries [41], we find that the distribution of terms in text-to-image prompts deviates from this pattern. Figure 1 shows that the frequencies of top-ranked terms present a milder decay than Zipf's law, and the tail terms present a clear exponential tail [8]. This is likely due to the specialized nature of creative activities, where the use of terms is more restricted than open Web search. This indicates the opportunity and feasibility of curating specialized vocabularies for creation, something similar to the Unified Medical Language System (UMLS) in the biomedical and health domain [2].

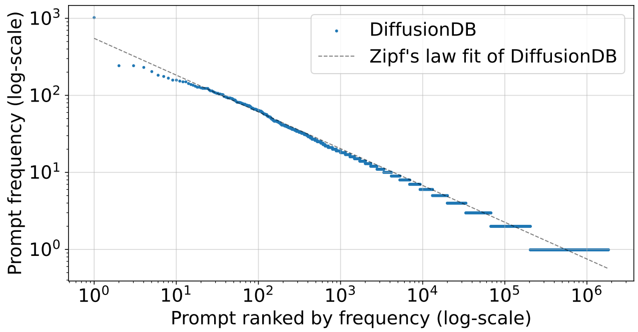

Figure 2: Prompt frequencies of DiffusinoDB plotted in a log-log scale. The distribution follows Zipf's law.

Table 5: Most frequent prompts in DiffusionDB. #Users indicates the number of users who have used this prompt.

Rank	Prompt	Freq.	#Users
1	painful pleasures by lynda benglis, octane render, colorful, 4k, 8k	1010	1
2	cinematic bust portrait of psychedelic robot from left, head and chest ...	240	2
3	divine chaos engine by karol bak, jean delville, william blake, gustav ...	240	7
4	divine chaos engine by karol bak and vincent van gogh	228	1
5	soft greek sculpture of intertwined bodies painted by james jean ...	202	2
6	detailed realistic beautiful young medieval queen face portrait ...	202	1
7	animation magic background game design with miss pokemon ...	181	2
8	cat	174	69
9	wrc rally car stylize, art gta 5 cover, official fanart behance hd ...	166	4
10	futurism movement hyperrealism 4k detail flat kinetic	157	1
11	a big pile of soft greek sculpture of intertwined bodies painted by ...	156	1
12	test	152	86
13	dream	149	133
14	realistic detailed face portrait of a beautiful futuristic viking warrior ...	149	2
15	spritesheet game asset vector art, smooth style beeple, by thomas ...	141	3
*16		137	50
17	abstract 3d female portrait age five by james jean and jason chan, ...	134	1
18	symmetry!! egyptian prince of technology, solid cube of light, ...	130	1
19	retrofuturistic portrait of a woman in astronaut helmet, smooth ...	127	1
20	astronaut holding a flag in an underwater desert. a submarine is ...	127	1

* Row 16 is an empty prompt.

4.2.2 *Prompt frequencies follow the power law.*

We also examine the distribution of prompt frequencies. From Figure 2, we find the prompt frequency distribution of the larger dataset, DiffusionDB, does follow Zipf's law (except for the very top-ranked prompts), similar to the queries of Web and vertical search engines [36, 41, 44]. The most frequently used prompts are listed in Table 5. Interestingly, many of the top-ranked prompts are (1) lengthy and (2) only used by a few users. This indicates that although the prompt frequency distributes are similar to that of Web search, the mechanism underneath may be different (shorter Web queries are more frequent and shared by more users [36]).

4.2.3 *Text-to-image generation prompts tend to be longer.*

We report the key statistics of prompt length (*i.e.*, the number of terms in a prompt) in Table 6. The average length of prompts for text-to-image generation (27.16 for Midjourney and 30.34 for DiffusionDB) and the median length (20 for Midjourney and 26 for DiffusionDB) are significantly longer than the lengths of Web search queries, where the mean is around 2.35 and the median is about 2 terms [14, 36]). Interestingly, similar observations are reported in vertical search engines such as electronic health records (EHR) search engines, where the queries are also significantly longer than

Table 6: Statistics of prompt lengths.

Dataset	Midjourney	DiffusionDB	SAC
Avg. #terms	27.16	30.34	17.53
Std. #terms	24.11	21.25	11.27
Median #terms	20	26	15
Max #terms	426	540	62

Web search queries (the average length is 5.0) [44], likely due to the highly specialized and complex nature of the tasks.

Bundled queries. When queries are more complex and harder to compose, an effective practice used in medical search engines is to allow users to *bundle* a long query, save it for reuse, and share it with others. In the context of EHR search, *bundled* queries are significantly longer (with 58.9 terms on average, compared to 1.7 terms in user typed-in queries) [44, 46]. Bundled queries tend to have higher quality, and once shared, are more likely to be adopted by other users [46]. Table 5 seems to suggest the same opportunity, as certain well-composed lengthy queries are revisited many times by their users. These prompts could be saved as "bundles" and potentially shared with other users. To illustrate the potential, we calculate the prompts used by multiple users and plot the distribution in Figure 3. We find a total of 16,950 unique prompts (0.94% of all unique prompts) have been used across users, 782 have been used by five or more users, and 182 have been shared by 10 or more users. The result suggests that text-to-image generation users have already started to share *bundled* prompts spontaneously, even though this functionality has not been provided by the system. Compared to vertical search engines that provide bundle-sharing features, the proportion of *bundled* prompts is still relatively small (compared with 19.3% for an EHR search engine [44]), indicating a huge opportunity for bundling and sharing prompts.

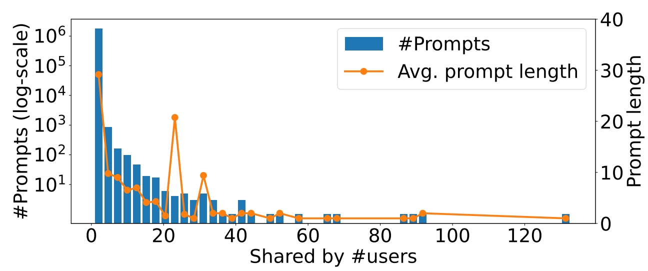

Figure 3: Prompts shared across users in DiffusionDB. The orange line plots the average prompt length in the blue bins.

4.2.4 *Text-to-image generation sessions contain more prompts.*

A session is defined as a sequence of queries made by the same user within a short time frame in Web search [44], which often corresponds to an atomic mission for a user to achieve a single information need [15, 36]. Analyzing sessions is critical in query log analysis because a session provides insights about how a user modifies the queries to fulfill the information need [15, 36].

Following the common practice in Web search, we chunk prompts into sessions with a 30-minute timeout [14, 44], meaning any two

Table 7: Prompts shared by the largest numbers of users in DiffusionDB. Only prompts longer than five terms are reported below row 10.

Rank	Prompt	#Users
1	dream	133
2	stable diffusion	91
3	help	89
4	test	86
5	cat	69
6	nothing	66
7	god	58
8	the backrooms	53
*9		50
10	among us	44
19	a man standing on top of a bridge over a city, cyberpunk art ...	32
20	mar - a - lago fbi raid lego set	32
34	an armchair in the shape of an avocado	23
35	a giant luxury cruiseliner spaceship, shaped like a yacht, ...	23
42	a portrait photo of a kangaroo wearing an orange hoodie and ...	19
45	anakin skywalker vacuuming the beach to remove sand	19
48	emma watson as an avocado chair	18
64	milkyway in a glass bottle, 4k, unreal engine, octane render	16

* Row 9 is an empty prompt.

consecutive prompts that are submitted by the same user within 30 minutes will be considered as in the same session.

The statistics of sessions are listed in Table 8. Similar to prompts, text-to-image generation sessions also tend to be significantly longer than Web search sessions (by the number of prompts in a session). A text-to-image generation session contains 10.25 or 13.71 (Midjourney or DiffusionDB) prompts on average and a median of 4 or 5 (Midjourney or DiffusionDB) prompts; while in Web search, the average session length is around 2.02 and the median is 1 [36]. This is again likely due to the complexity of the creation task so the users need to update the prompts multiple times. Indeed, a user tends to change (*add*, *delete*, or *replace*) a median of 3 terms (measured by term-level edit distance) between two consecutive prompts in the same session on Midjourney (5 on DiffusionDB), astonishingly more than how people update Web search queries. Do these updates indicate different types of information needs?

Table 8: Statistics of prompt sessions. Sessions are identified with a 30-minute timeout. Edit distances regarding terms are calculated with consecutive prompts in the same session.

Dataset	Midjourney	DiffusionDB
#Sessions	14,232	161,001
Avg. #sessions/user	8.52	15.51
Median #sessions/user	2	9
Avg. #prompts/session	10.19	13.71
Median #prompts/session	4	5
Avg. edit distance	8.53	9.42
Median edit distance	3	5

4.2.5 ***A new categorization of information needs.*** Web search queries are typically distinguished into three categories: (1) *navigational queries*, (2) *informational queries*, and (3) *transactional queries* [3]. Should text-to-image prompts be categorized in the same way? Or do prompts express new categories of information needs?

Navigational prompts. The most frequent queries in Web search are often *navigational*, where users simply use a query to lead them to a particular, known Website (*e.g.*, "Facebook" or "YouTube"). In text-to-image generation, as the generation model often returns different images given the same text prompt due to randomization, the information need of "navigating" to a known image is rare. Indeed, the queries used by the most number of users (Figure 3) are generally not tied to a particular image. Even though the shorter queries on the top look somewhat similar to "Facebook" or "Youtube", are rather ambiguous and more like testing the system.

Informational prompts. Most other text-to-image prompts can be compared to *informational* queries in Web search, which aim to acquire certain information that is expected to present on one or more Web pages [3]. The difference is that informational prompts aim to *synthesize* (rather than *retrieve*) an image, which is expected to exist in the *latent* representation space of images. Most prompts fall into this category, similar to the case in Web search [3].

Transactional prompts. Transactional queries are those intended of performing certain Web-related activities [3], such as completing a transaction (*e.g.*, to book a flight or to make a purchase). One could superficially categorize all prompts into transactional, as they are all intended to conduct the activities of "generating images". Zooming into this superficial categorization, we could identify prompts that refer to specific and recurring tasks, such as "3D rendering", "post-processing", "global illumination", and "movie poster" (see more examples in Section 4.1.2). These tasks may be considered transactional in the context of text-to-image generation.

Exploratory prompts. Beyond the above categories corresponding to the three basic types of Web search queries, we discover a new type of information needs in prompts, namely the *exploratory* prompts for text-to-image generation. Comparing to an *informational* prompt that aims to generate a specific piece of (hypothetical) image, an *exploratory* prompt often describes a vague or uncertain information need (or image generation requirements) that intentionally leads to multiple possible answers. The user intends to explore different possibilities, leveraging either the randomness of the model or the flexibility of terms used in a prompt session.

Indeed, rather than clearly specifying the requirements and constraints and gradually refining the requirements in a session, in *exploratory* prompts or sessions, the users tend to play with alternative terms of the same category (*e.g.*, different colors or animals, or *sibling* terms) to explore how the generation results could be different or could cover a broader search space. Based on the session analysis, we count the most frequent term replacements in Table 9. In this table, we find 36 replacements that show *exploratory* patterns, such as ("man", "woman"), ("cat", "dog"), ("red", "blue"), and ("16:9", "9:16").

On the contrary, in *non-exploratory* sessions, replacing a term with its synonyms or hyponyms, or more specific concepts are more common, which refines the search space (rather than exploring the generation space). In the table, we find a few such replacements such as ("insect", "ladybug") and ("painting", "portrait"). There are also examples that replace terms with the correct spelling or replace punctuations to refine: ("aphrodesiac", "aphrodisiac"), ("with", ","), (",", "and") and (",", ".").

Table 9: Most frequent term replacements. This table only considers consecutive prompts from the same session where exactly one term is been replaced. Green highlights replacements that might indicates *exploratory* patterns, while red highlights *non-exploratory* replacements.

	Midjourney		DiffusionDB	
	Replacement	Freq.	Replacement	Freq.
1	(deco, nouveau)	16	(man, woman)	216
2	(16:9, 9:16)	15	(woman, man)	187
3	(9:16, 16:9)	14	(2, 3)	161
4	(2, 1)	8	(1, 2)	147
5	(16:9, 4:6)	8	(7, 8)	140
6	(1, 2)	7	(8, 9)	139
7	(3:4, 4:3)	7	(6, 7)	135
8	(1000, 10000)	7	(3, 4)	132
9	(artwork, parrot)	7	(girl, woman)	128
10	(16:9, 1:2)	6	(red, blue)	116
11	(2:3, 3:2)	6	(5, 6)	115
12	(asian, white)	6	(4, 5)	112
13	(1, 0.5)	5	(female, male)	107
14	(320, 384)	5	(male, female)	97
15	(0.5, 1)	4	(blue, red)	93
16	(crown, throne)	4	(0, 1)	89
17	(blue, green)	4	(cat, dog)	89
18	(9:16, 4:5)	4	(woman, girl)	82
19	(2:3, 1:2)	4	(dog, cat)	79
20	(--w, --h)	4	(white, black)	72
21	(nouveau, deco)	4	(with, ",")	71
22	(red, blue)	4	(steampunk, cyberpunk)	71
23	(guy, girl)	4	(red, green)	70
24	(snake, apple)	4	(cyberpunk, steampunk)	70
25	(japanese, korean)	4	(",", and)	69
26	(16:8, 8:11)	4	(painting, portrait)	68
27	(insect, ladybug)	4	(",", ".")	68
28	(--hd, --vibe)	3	(portrait, painting)	68
29	(aphrodesiac, aphrodisiac)	3	(girl, boy)	64
30	(0.5, 2)	3	(green, blue)	63

Another indication of *exploratory* behavior is the repeated use of prompts. For example, among the top prompts in Table 5 (except those for testing purposes), each of them is repeatedly used by the same user more than 100 times. This might be because the user is exploring different generation results with the same prompt, leveraging the randomness of the generative model.

4.3 How are the Information Needs Satisfied?

Prompts are typically crafted to meet certain information needs by generating satisfactory images. In this subsection, we examine how prompts can fulfill this goal. With the rating annotations in the SAC dataset (the average rating is 5.53, and the median is 6), we calculate the correlation between ratings and other variables like prompt lengths and term frequencies.

4.3.1 ***Longer prompts tend to be higher rated.*** We plot how the ratings of generated images correlate with prompt lengths in Figure 4, where we find a positive correlation with the Pearson coefficient at 0.197. This means longer prompts tend to produce images of higher quality. This provides another perspective to understand the large lengths of prompts and prompt sessions and another motivation to bundle and share long prompts.

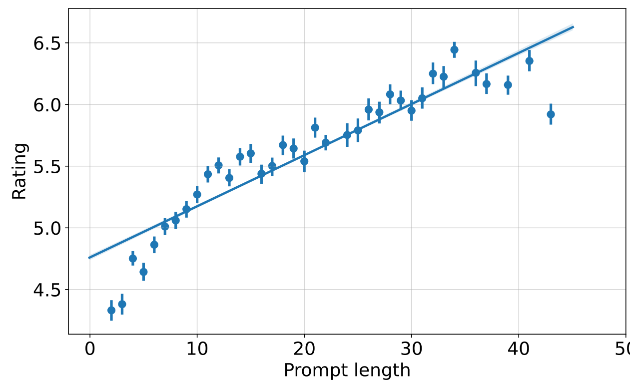

Figure 4: Prompt length is positively correlated with ratings. The Pearson correlation coefficient is 0.197.

4.3.2 ***The choice of words matters.*** We also investigate how the choice of words influences the performance of image generation. We collect all the prompts that contain a particular term and calculate the average rating. Terms with the highest and lowest average ratings are listed in Table 12 in the appendix. We find most high-rating terms are artist names, which provide clear constraints on the styles of images. In contrast, terms with low ratings are much vaguer and more abstract and might indicate an *exploratory* behavior. More efforts needed to be done to handle *exploratory* prompts and to encourage the users to refine their needs.

4.4 How are Users' Information Needs Covered by Image Captions?

Current text-to-image generation models are generally trained with large-scale image-text datasets, where the paired text usually come from image captions. To figure out how these training sets match the actual users' information needs, we compare the prompts with image captions in the open domain. In particular, we consider LAION-400M [35] as one of the main sources of text-to-image training data since both LDMs and the Stable Diffusion model employ this dataset. Text in LAION-400M are extracted from the captions of the images collected from the Common Crawl, so they are supposed to convey the subject, form, and intent of the images. We randomly sample 1M texts from LAION-400M and compare them with user-input prompts. We obtain the following finding.

Term usages are different between user-input prompts and image captions in open datasets. We construct a vocabulary based on LAION-400M and calculate the vocabulary coverage of three prompt datasets (*i.e.*, to what proportion of the user-input terms is covered by the LAION vocabulary). The coverage is 25.94% for Midjourney, 43.17% for DiffusionDB, and 80.56% for SAC. The coverage is relatively high on SAC as this dataset is relatively clean. In comparison, the Midjourney and DifussionDB datasets directly collect prompts from Discord channels of Midjourney and Stable Diffusion, and over half of the terms are not covered in the LAION dataset. We also analyzed their embeddings and find that user-input prompts and image captions from the LAION dataset cover very different regions in the latent space (Figure 8 in the appendix).

5 IMPLICATIONS

Our analysis presents unique characteristics of user-input prompts, which helps us better understand the limitations and opportunities of text-to-image generation systems and AI-facilitated creativity on the Web. Below we discuss a few concrete and actionable possibilities for improving the generation systems and enhancing creativity.

Building art creativity glossaries. As we discussed in Sec. 4.1.1, a text-to-image prompt could be decomposed into three aspects: *subject* ("what"), *form* ("how"), and *intent* ("why", or *content* as in classical Art literature). If we can identify and analyze these specific elements in prompts, we may be able to better decipher users' information needs.

However, to the best of our knowledge, there is no existing tool that is able to extract the *subject*, *form*, and *intent* from text prompts. Besides, although users have spontaneously collected terms that describe the *form* and *subject*[8], there is no high-quality and comprehensive glossary in the literature that contains terms about these three basic components of art, or something like the Unified Medical Language System (UMLS) for biomedical and health domains [2]. Constructing such tools or glossaries is difficult and will highly rely on the domain knowledge, because: (1) These three components of art are often intertwined and inseparable in a piece of work [22], meaning a term would have tendencies to fall into any categories of these three. For example, in *Process Art*, the *form* and *content* seem to be the same thing [22]. (2) Terminologies about art are consistently updated because new artists, styles, and art-related sites keep popping out. We call for the joint effort of the art and the Web communities to build such vocabularies and tools.

Bundling and sharing prompts. Sec. 4.2.3 analyzes the lengths of text-to-image prompts, where we find an inadequate use of *bundled* prompts compared with other vertical search engines (*e.g.*, EHR search engines). Since the prompts are generally much longer than Web search queries, and the information needs are also more complex, it is highly likely that *bundled* prompts can help the users to craft their prompts more effectively and efficiently. Though there are already prompt search websites like Lexica[9], PromptHero[10] and PromptBase[11] that provide millions of user-crafted prompts, such *bundled* search features are merely integrated into current text-to-image generation systems. As mentioned earlier, adding features to support bundling and sharing high-quality prompts could bring immediate benefits to text-to-image generation systems.

Personalized generation. The analysis in Sec. 4.2.4 suggests that the session lengths in text-to-image generation are also significantly larger than the session lengths in Web search, indicating the great opportunity for a personalized generation. Currently, the *session-based* generation features are mostly built upon image initialization of diffusion models, *i.e.*, using the output from the previous generation as the starting point of diffusion sampling. Compared with other *session-based* AI systems like ChatGPT [23], these *session-based* features still seem preliminary and take little consideration about personalized generation. Meanwhile, the explicit descriptions of *forms* and *intent* in prompts also indicate opportunities to customize the generation models for these constraints (and the potential applications as listed in Section 4.1.2).

Handling exploratory prompts and sessions. In Sec. 4.2.5 we identify a new type of prompt in addition to the three typical categories of query in Web search (*i.e.*, *navigational*, *informational*, and *transactional* queries), namely the *exploratory* prompts. To encourage the *exploratory* generation of images, reliable and informative *exploration measures* will be much needed. In other machine innovation areas, like AI for molecular generation, efforts have been made on discussing the measurement of coverage and exploration of spaces [42, 43], but for text-to-image generation, such discussions are still rare. How to encourage the models to explore a larger space, generate novel and diverse images, and recommend exploratory prompts to users are all promising yet challenging directions.

Improving generation models with prompt logs. Finally, the gap between the image captions in open datasets and the user-input prompts (Sec. 4.4) indicates that it is desirable to improve model training directly using the prompt logs. Following the common practice in Web search engines, one may leverage both explicit and implicit feedback from the prompt logs (such as the ratings or certain behavioral patterns or modifications in the prompts) as additional signals to update the generation models.

Although we focus our analysis on text-to-image generation, the analogy to Web search and some of the above implications also apply to other domains of AI-generated content (AIGC), such as AI chatbots (*e.g.*, ChatGPT).

6 CONCLUSION

We take an initial step to investigate the information needs of text-to-image generation through a comprehensive and large-scale analysis of user-input prompts (analogous to Web search queries) in multiple popular systems. The results suggest that (1) text-to-image prompts are typically structured with terms that describe the *subject*, *form*, and *intent*; (2) text-to-image prompts are sufficiently different from Web search queries. Our findings include the significantly lengthier prompts and sessions, the lack of *navigational* prompts, the new perspective of *transactional* prompts, and the prevalence of *exploratory* prompts; (3) image generation quality is correlated with the length of the prompt as well as the usage of terms; and (4) there is a considerable gap between the user-input prompts and the image captions used to train the models. Based on these findings, we present actionable insights to improve text-to-image generation systems. We anticipate our study could help the text-to-image generation community to better understand and facilitate creativity on the Web.

[8]Prompt book for data lovers II: https://docs.google.com/presentation/d/1V8d6TlIKqB1j5xPFH7cCmgKOV_fMs4Cb4dwgjD5GIsg, retrieved on 3/14/2023.
[9]Lexica: https://lexica.art/, retrieved on 3/14/2023.
[10]PromptHero: https://prompthero.com/, retrieved on 3/14/2023.
[11]PromptBase: https://promptbase.com/, retrieved on 3/14/2023.

REFERENCES

[1] Alan Agresti. 2012. *Categorical data analysis*. Vol. 792. John Wiley & Sons.

[2] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[3] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.

[4] 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.

[5] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *EMNLP*.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[8] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[9] Thomas H. Davenport and Nitin Mittal. 2022. How generative AI is changing creative work. https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work Retrieved on 3/15/2023..

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc.

[12] Jorge R Herskovic, Len Y Tanaka, William Hersh, and Elmer V Bernstam. 2007. A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association* 14, 2 (2007), 212–220.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[14] Bernard J Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. 1998. Real life information retrieval: A study of user queries on the web. In *ACM Sigir Forum*, Vol. 32. ACM New York, NY, USA, 5–17.

[15] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 699–708.

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] 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.

[18] Elman Mansimov, Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. 2016. Generating Images from Captions with Attention. In *ICLR*.

[19] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[20] Midjourney.com. 2022. Midjourney. https://midjourney.com/ Retrieved on 3/15/2023..

[21] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. PMLR, 16784–16804.

[22] Otto G Ocvirk, Robert E Stinson, Philip R Wigg, Robert O Bone, and David L Cayton. 1968. *Art fundamentals: Theory and practice*. WC Brown Company.

[23] OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt Retrieved on 3/15/2023..

[24] OpenAI. 2023. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf Retrieved on 3/15/2023..

[25] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference*. 192–202.

[26] Jonas Oppenlaender. 2022. A Taxonomy of Prompt Modifiers for Text-to-Image Generation. *arXiv preprint arXiv:2204.13988* (2022).

[27] Nikita Pavlichenko and Dmitry Ustalov. 2022. Best Prompts for Text-to-Image Models and How to Find Them. (2022).

[28] John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. 2022. Simulacra Aesthetic Captions. https://github.com/JD-P/simulacra-aesthetic-captions Retrieved on 3/15/2023..

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[32] 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.

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*.

[34] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv preprint arXiv:2301.09515* (2023).

[35] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

[36] 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.

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[39] Iulia Turc and Gaurav Nemade. 2022. Midjourney User Prompts & Generated Images (250k). https://doi.org/10.34740/KAGGLE/DS/2349267 Retrieved on 3/15/2023..

[40] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. *arXiv preprint arXiv:2210.14896* (2022).

[41] Yinglian Xie and David O'Hallaron. 2002. Locality in search engine queries and its implications for caching. In *Proceedings. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, Vol. 3. IEEE, 1238–1247.

[42] Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. 2022. How Much of the Chemical Space Has Been Explored? Selecting the Right Exploration Measure for Drug Discovery. In *ICML 2022 2nd AI for Science Workshop*.

[43] Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. 2023. How Much Space Has Been Explored? Measuring the Chemical Space Covered by Databases and Machine-Generated Molecules. In *The Eleventh International Conference on Learning Representations*.

[44] Lei Yang, Qiaozhu Mei, Kai Zheng, and David A Hanauer. 2011. Query log analysis of an electronic health record search engine. In *AMIA annual symposium proceedings*, Vol. 2011. American Medical Informatics Association, 915.

[45] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research* (2022).

[46] Kai Zheng, Qiaozhu Mei, and David A Hanauer. 2011. Collaborative search in electronic health records. *Journal of the American Medical Informatics Association* 18, 3 (2011), 282–291.

A DATA AND DATA PROCESSING

A.1 Datasets

For the datasets, we list important features (prompt, timestamp, user ID, and rating), feature descriptions, and corresponding examples in Table 10.

Table 10: Feature descriptions and examples of the Midjourney, DiffusionDB, and SAC datasets.

	Description	Examples
Prompt	The prompt used to generate images. **Type**: `String`	**Midjourney**: `hands, by Karel Thole and Mike Mignola --ar 2:3` **DiffusionDB**: `Ibai Berto Romero as Willy Wonka, highly detailed, oil on canvas` **SAC**: `concept art by David Production.`
Timestamp	The timestamp when the image was generated from the prompt. **Type**: `String`	**Midjourney**: `2022-06-23T23:58:16.024000+00:00` **DiffusionDB**: `2022-08-07 22:57:00+00:00` **SAC**: N/A
User ID	The unique ID for the user account who submitted the prompt. **Type**: `String`	**Midjourney**: `977252506335858758` **DiffusionDB**: `fcdb3e09f977412c342b6624a19d1295ee1334c153c90af16d1cca8d9f27b04a` **SAC**: N/A
Ratings	The rating of the image generated from the prompt[12]. **Type**: `Integer`	**Midjourney**: N/A **DiffusionDB**: N/A **SAC**: 6, 5, 7, 5

A.2 Data Processing

Midjourney. We extracted prompts, timestamps, and user IDs from the records in the Midjourney dataset. The prompts in Midjourney may contain specific syntactic parameters of the Midjourney model, such as "`--ar`" for aspect ratios, "`--h`" for heights, "`--w`" for widths, "`::`" for assigning weights to certain terms in the prompts. We first take the lowercase characters from tokenized prompts with the Spacy tokenizer[13]. Regarding the parameters, such as "`--h`", we consider them single terms. Specially, we split the weighted terms with their weights, and consider "`::`" and "`::-`" (negative weight) as two different terms. During tokenization, we also removed redundant whitespaces. Midjourney allows users to upload reference images as parts of their prompts in the form of Discord links. These links are also processed as special terms.

DiffusionDB. We utilize the metadata of DiffusionDB-Large (14M) for prompt analysis. We first remove duplicate data entries with the same prompt, timestamp, and user ID, meaning these entries record different images generated by the same user with the same prompt as a single submission. As a result, we obtained 2,208,019 non-duplication prompt submissions from users. Note that repeated submissions of prompts are reserved. We tokenize the prompts and remove the whitespace as we process the Midjourney data.

SAC. SAC provides aesthetic ratings of generated images. Note that one prompt can correspond to multiple images, and each image can also have multiple ratings. Since there are no user ID and timestamp annotations in SAC, to remove the duplicates, we simply extract the unique prompts and conclude all correlated ratings.

More details can be found in the supplementary materials.

B ADDITIONAL ANALYSIS RESULTS

B.1 Prompt-Level Analysis

Prompt length distributions. The distributions of prompt lengths are displayed in Figure 5, where the modes are around 10.

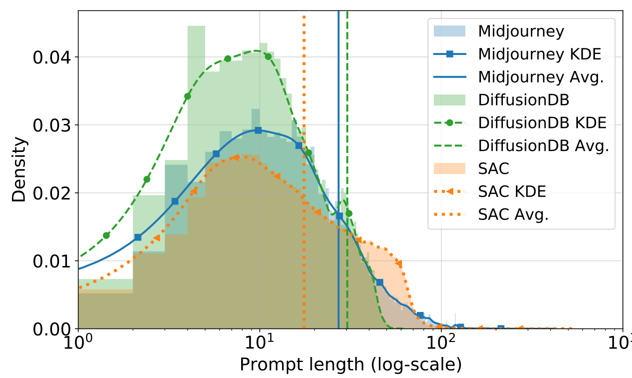

Figure 5: Prompt length distributions. The x-axis (prompt length) is plotted in the log scale.

Prompts revised by users. Table 11 lists the most revisited prompts in DiffusionDB.

Time series analysis. We analyze how the prompts distribute within 24 hours for the Midjourney and DiffusionDB datasets. The results are shown in Figure 6. The patterns in these two datasets are similar: the rushing hours are around 01:00–03:00 (for both Midjourney and DiffusionDB), 15:00–17:00 (Midjourney), 20:00 (DiffusionDB); while during the daytime, the users are relatively inactive.

Ratings. The overall rating distribution of SAC is displayed in Figure 7.

B.2 Comparing Prompts with Training Data

To compare user-input prompts with texts that are used to train the text-to-image generation models, we also include the LAION dataset [35]. LAION is a public dataset of CLIP-filtered image-text pairs and has often been used in large text-to-image model training [32–34, 45]. In the analysis, we use the LAION-400M dataset[14] that contains only English texts.

[12] Note that one prompt may correspond to multiple images, and one image may have multiple ratings. Here we list all the ratings correlated to the example prompt.
[13] Spacy: https://spacy.io/, retrieved on 3/15/2023.
[14] LAION-400M dataset: https://laion.ai/blog/laion-400-open-dataset/.

Table 11: Most revisited prompts in DiffusionDB. Only revisits across sessions are considered.

	Prompt	#Revisits
1	test	24
2	cat	19
3	fat chuck is mad	15
4		15
5	dog	15
6	symmetry!! egyptian prince of technology, solid cube of light, ...	13
7	full character of a samurai, character design, painting by gaston ...	13
8	studio portrait of lawful good colorful female holy mecha paladin ...	11
9	full portrait and/or landscape. contemporary art print. high taste. ...	11
10	woman wearing oculus and digital glitch head edward hopper and ...	11
11	dream	10
12	hyperrealistic portrait of a character in a scenic environment by ...	10
13	full portrait &/or landscape painting for a wall. contemporary art ...	10
14	zombie girl kawaii, trippy landscape, pop surrealism	10
15	creepy ventriloquist dummy in the style of roger ballen, 4k, bw, ...	9
16	cinematic bust portrait of psychedelic robot from left, head and ...	9
17	red ball	9
18	amazing landscape photo of mountains with lake in sunset by ...	9
19	female geisha girl, beautiful face, rule of thirds, intricate outfit, ...	9
20	full portrait and/or landscape painting for a wall. contemporary ...	9

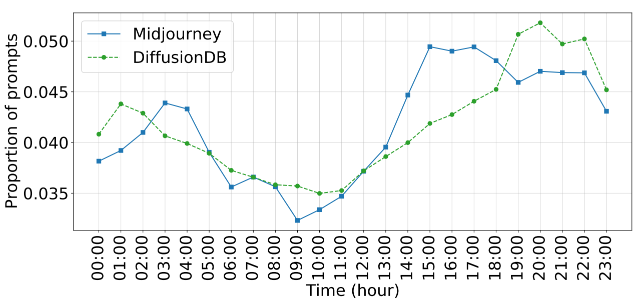

Figure 6: The distribution of prompts within 24 hours.

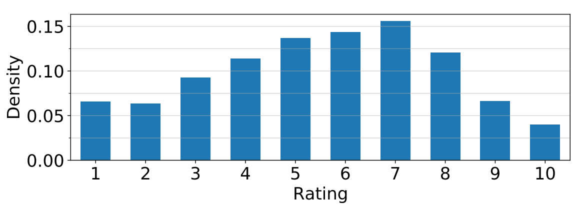

Figure 7: Rating distribution. Each rating corresponds to a user-input prompt and an image generated from that prompt. The average rating is 5.53, the standard deviation is 2.40, and the median is 6.

Visualization. To intuitively see how user-input prompts and texts from the LAION training set are distributed, we use UMAP [19] to visualize the prompts and the texts based on BERT [10] embeddings in Figure 8. In the visualization, we find a clear gap between LAION (red circles) and other datasets, meaning the training set can hardly represent the real data distributions of user-input prompts. This visualization also aligns with the findings about vocabulary coverage, where we discover the terms in SAC are most covered by LAION, and the vocabulary of Midjorney is most distant from that of LAION.

Table 12: Terms with the highest and the lowest average ratings. Only terms with frequencies larger than 100 are considered. "Avg." and "Std." are means and standard deviations of ratings respectively.

	Terms with highest avg. ratings				Terms with lowest avg. ratings			
	Term	Avg.	Std.	Freq.	Term	Avg.	Std.	Freq.
1	shinjuku	8.55	0.90	168	equations	2.36	2.18	240
2	gyuri	8.22	1.65	219	mathematical	2.37	2.18	230
3	lohuller	8.22	1.66	215	geismar	2.67	2.13	136
4	afremov	7.95	1.73	288	haviv	2.68	2.14	136
5	leonid	7.95	1.73	288	chermayeff	2.73	2.14	136
6	retrofuture	7.95	1.97	307	learning	3.10	2.64	112
7	merantz	7.93	1.77	463	pegasus	3.10	2.00	129
8	josan	7.91	1.73	1,647	teacher	3.11	2.59	110
9	fantasyland	7.90	1.52	114	someone	3.14	2.45	574
10	gensokyo	7.89	1.34	281	funny	3.17	2.52	208

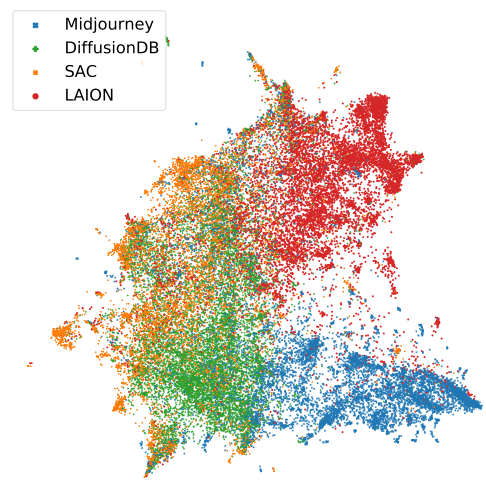

Figure 8: UMAP visualization of prompt embeddings. A clear gap can be identified between the LAION training data (red circles) and the user-input prompts (other colors).

Non-representative training data. We discover a huge gap between the user-input prompts and the texts in the open training data such as the LAION training set. The out-of-vocabulary (OOV) problem is severe, and in the prompts from Midjorney, about 75% terms are not covered by LAION's vocabulary. Figure 8 also displays a gap in prompt embedding distributions. All this evidence proves that the texts (mostly image captions) from the open training data can hardly represent users' information needs and we should call for another way that renders better supervision during training. ChatGPT [23] has already demonstrated that reinforcement learning from human feedback (RLHF) [7] could provide rich supervision and guidance to the model. However, for text-to-image generation, related work is still limited. Note that our analysis is based on the open datasets that are included in the training data of the models and doesn't consider the private training data that could have a different coverage of the space.

Received 7 February 2023; revised 15 March 2023; accepted 6 March 2023