# Examining the Text-to-Image Community of Practice: Why and How do People Prompt Generative AIs?

Téo Sanchez

# Examining the Text-to-Image Community of Practice: Why and How do People Prompt Generative AIs?

TÉO SANCHEZ, Selas Studio, France

Image generation gained popularity with machine learning (ML) models generating images from text, fuelling new online communities of practices. This work explores the sociology, motivations, and usages of AI art hobbyists. We analyzed an online questionnaire answered by 64 practitioners and a dataset of user prompts sent to the Stable Diffusion generative model. Our findings suggest that TTI generation is a recreational activity mainly conducted by narrow socio-demographic groups who use auxiliary techniques across platforms and beyond request-response interactions. Inherent model limitations and finding suitable prompt formulation are the main obstacles practitioners face. A taxonomy and a corresponding ML model capable of recognizing the semantic content of unseen prompts were created to conduct the user prompt analysis. The prompt analysis revealed that artist names are the main specifier used beside the main subject, often in sequences. We finally discuss the design and socio-technical implications of our work for creativity support.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Information extraction.*

Additional Key Words and Phrases: text-to-image generation, community of practice

## 1 INTRODUCTION

In 2022, image generation experienced significant advancements as generative ML models leveraged larger and more diverse datasets to create coherent and high-resolution images. More importantly, generative models became controllable from natural language text entries, called prompts, guiding the desired characteristic of the image generation as illustrated in Figure 1.

```
An old groundhog wearing a cloak and an ivy
cap, outdoor photography, distant mountains, animal
documentary, cinematic, movie by Denis Villeneuve
```

**(a)**

**(b)**

**Fig. 1. An example of text-to-image prompt (a) and its associated image generation (b).**

These technological advancements transformed generative AI from a niche artistic pursuit to a widespread and recreational online practice. Indeed, private companies brought TTI generation to the mainstream, although adopting different policies: Midjourney [6] provided access to its TTI models through a Discord server, emphasizing community building. In contrast, Stability AI made their TTI model publicly available and offered an API, fuelling the proliferation of commercial text-to-image applications. With this new mainstream access to TTI generative models, users organized on social media to share their creations and tips about how to write text prompts to generative AI models. Understanding this recent and fast-evolving community of practice is essential as the public debate on its economic and ethical consequences becomes increasingly polarizing. Furthermore, understanding this emerging practice can inform the design of meaningful interactions with generative ML systems.

*C&C, June 19–21, 2023, Virtual*

Existing HCI research derived guidelines for writing effective prompts using various methods, such as controlled experiments with few human participants [34], using algorithmic techniques [21, 35, 41], and auto-ethnographic approach to categorize prompts [39]. However, the usages and motivations behind TTI generation have yet to be investigated from the perspective of actual generative AI hobbyists.

The current study employs a bottom-up ethnographic approach, combining an online questionnaire and an analysis of a large collection of user-generated prompts. It aims to address the following research questions:

**RQ1** What are the motivations and socio-demographic factors of TTI practitioners?
**RQ2** What are the challenges associated with the use of TTI generators?
**RQ3** What are the usage patterns of TTI practitioners? In particular, how to analyze the semantic structure of users' TTI prompts?

Based on insights gained, this research aims to identify guidelines for designing better interactions with TTI generators and, more generally, deepening our understanding of the TTI generation community of practice, ultimately contributing to a more informed grasp of the public debate around generative AI.

After providing some background and existing literature on text-to-image prompt engineering, the paper reports on the results of the online questionnaire disseminated among AI art online communities to progress toward the three research questions. Section 5 describes a methodology for analyzing a large corpus of text-to-image prompts and identifying higher-level patterns in prompt structures. The training of a named entity recognition (NER) model used to conduct this analysis is described and made available. Section 6 discusses implications for designing meaningful interactions with text-based synthetic media generators that could foster the discoverability of new artistic jargon or contribute to recognizing artists invoked in text-to-image prompts. The main contributions of this research are:

- Uncovering socio-demographic factors of online TTI communities and their motivations for using TTI generators;
- Identifying the usage challenges with TTI generative systems;
- Uncovering structural patterns of TTI prompts by training and publishing a NER model to analyze a large corpus of prompts;
- Provide design implications based on both our empirical results and the technological block developed to analyze TTI prompts.

## 2 BACKGROUND AND RELATED WORK

In this section, we provide a background on the evolution of AI art and define TTI generation as a sub-field of AI art. We then present existing HCI research on text-to-image prompting practices and systems, emphasizing the links with the existing literature on creativity-support tools. Given the scarcity of published work on this recent topic, we include peer-reviewed and non-peer-reviewed references mainly published on arXiv [23].

### 2.1 Brief history of AI art

Artificial intelligence (AI) art has been evolving as a form of contemporary media and digital art since the mid-2010s. In the early days of AI art, a limited number of artists utilized data and machine learning (ML) as artistic material, in conjunction with advancements in model architecture from research. The Generative Adversarial Network (GAN) was a particularly influential model in AI visual art from 2015 to 2020. The ability for GANs to be trained on self-curated datasets using a single GPU permitted independent artists to gain autonomy and ownership of this technology.

GANs introduce unique artifacts and a recognizable aesthetic ironically referred to as "GANism" by artists and scholars alike [37]. The study conducted by Caramiaux and Alaoui [19], which interviewed five early pioneering AI artists, highlighted the political significance of their work in response to the monopolistic deployment of AI by industry. This finding aligns with previous research [25, 52] that also discussed the critical stance held by early AI artists against the cultural and political normalization induced by science and technology.

In the early 2020s, several ML advances radically changed AI art from a niche artistic pursuit to a social phenomenon. First, the use of CLIP[43], a pretrained multi-modal model developed by OpenAI, permitted the general public to generate images using natural language. Second, the quality and coherence of generated images were improved by diffusion models [28, 29, 49, 50]. Third, these models were trained on an immense number of text-image pairs collected from the web by non-profit research-oriented organizations [1, 47], which greatly expanded the variability of concepts to be explored with text-to-image generation. These achievements made the exploration of text-to-image generators limitless and accessible to the public. Small online communities first adopted text-to-image generation before it gained widespread media attention and became popular among millions of users by the end of 2022.

This article focuses on the latest form of AI art that emerged in the early 2020s, which we will refer to as text-to-image (TTI) generation. Unlike the early AI art movement, TTI generation has distanced itself from the contemporary art realm to become an online phenomenon[1] with significant social, economic, and legal consequences for the creative and cultural industries.

### 2.2 Understanding text-to-image prompting

The immense volume of text prompts addressed to TTI generators in just a few months offers a unique opportunity to understand the ways humans engage in creative communication with AI systems. Practice-oriented research on this topic is still early, but existing studies have contributed valuable insights into this practice. Liu and Chilton [34] examined the impact of various prompt specifiers and model hyperparameters (i.e., generation parameters besides text entry) on the coherence of text-to-image model outputs. They conducted five experiments in which two knowledgeable creatives evaluated the quality of thousands of images generated by combining prompt specifiers. Their analysis led to the development of seven guidelines to help users effectively interact with text-to-image models and produce improved results.

Other research took an algorithmic-centered approach to prompt crafting. Deckers et al. [21] reframed the prompt crafting process as an interactive image retrieval problem on infinite indexes, in which a prompt corresponds to a query. They discuss prospects for query-based workflow with generative models and highlight challenges and opportunities for IR research posed by the concept of an infinite index. Martins et al. [35] proposed MetaPrompter, an interactive tool to generate TTI prompt candidates using an evolutionary approach from an initial prompt blueprint. Pavlichenko and Ustalov [41] also sought to improve the quality of text-to-image model outputs by employing genetic algorithms to generate prompts combining different specifiers. Their results also suggest that using the most popular specifiers does not correlate with popular image results. Algorithmic approaches hold prospects for new workflows that would distance from typing prompts but rather generate candidates to be selected by users.

Oppenlaender [40] discussed the creativity associated with prompt crafting using James Rhodes's "the four P" conceptual model (person, process, press, product). The author argued that more than the product-centered view of creativity is needed for evaluating the creativity of text-to-image

[1] In-person social events around TTI generation exists, such as prompt battles [11], but remain too marginal to be considered in our focus.

generation and emphasizes the community nature of the practice. In a separate study, Oppenlaender [39] constructed a taxonomy of prompts using an auto-ethnographic approach, collecting and organizing prompts encountered online primarily via Twitter. However, data availability was limited until the publication of DiffusionDB [57], an open-source dataset of real users' prompt-image pairs addressed to the Stable Diffusion V1 model. The taxonomy was designed to support the activity of prompt engineering but does not provide implications for interactive creativity-support tools. Instead, it surveyed all prompting practices, including marginal usages such as specifier repetitions or "magical terms" i.e., abstract specifiers inducing unpredictability in the generation such as *"control the soul"* [39].

The present study adopts a bottom-up ethnographic approach to investigate the online TTI generation community. The methods used include analysis of both questionnaire data and 1.8 million unique user prompts from the DiffusionDB dataset [57].

### 2.3 Supporting text-to-image prompting

TTI prompting implies a task-divided co-creativity [31] where humans have the roles of task-definer and evaluator, while the system is a generator of concepts. Several tools and techniques for assisting and extending TTI were developed: inpainting and outpainting [2, 3], image prompts [4], face restoration algorithms [56], image to text [42], and model finetuning such as Dreambooth [45]. These techniques are dispersed among platforms, public computational notebooks, and API providers. Few approaches support the prompting process in itself i.e., finding the right words to express intentions and explore new possibilities. Three main approaches are currently available. First, auto-completion approaches comprise finetuned GPT-2 models published on the HuggingFace repository [46, 54]. Second, users can reuse real user prompts collected on large prompts repositories [5, 12] or markets [14]. Third, commercial applications provide dashboard interfaces for selecting specifier presets to construct prompts [7–10, 13]. These techniques might tend to homogenize practices rather than foster the discoverability of new vocabulary and have little or no adaptation to the users' intentions.

We foresee an HCI opportunity to design interactions that can onboard novice users while facilitating the discovery of new artistic jargon i.e., words or phrases that are unique to the art field such as "chiaroscuro", "trompe l'oeil", or "sfumato". This opportunity aligns with the design implications of Dang et al. [20] to formulate, combine, apply, and represent prompts in user interfaces to control large language models (LLM) for text-to-text generation. The authors highlight the importance of recognizing the semantic content of the prompt, such as task and style, to provide relevant user guidance. Our work adopts this approach and offers a theoretical and technological contribution to this end: a taxonomy and an associated ML model for recognizing the semantic content of unseen prompts.

## 3 DEFINITION AND ASSUMPTIONS ABOUT TEXT-TO-IMAGE PROMPTING

A typical text-to-image generative model first maps a sentence into a lower-dimensional vector representing visual concepts. The text encoder performs this mapping and the resulting latent representation then guides the generation of an image starting from a random seed vector. In practice, users explore what the model has learned by appending sentence fragments and keywords, steering the latent conditioning representation across various dimensions. Text-to-image prompting goes beyond simply appending keywords, as the model can capture semantic relationships between and within specifiers in the input text. Despite the widespread practice of separating specifiers with commas, most text encoder tokenizers (e.g. the CLIP tokenizer) do not consider these commas when processing the prompt. Therefore, punctuation is only a means for human users to structure the text prompts but has no actual impact on the system.

The following definitions and assumptions will be used throughout the article and analysis:

(1) We define **prompt specifiers** as text fragments specifying the desired characteristic of the image output. There is no definitive or deterministic way to split the prompt into specifiers, as commas do not necessarily separate them. Text-to-image practitioners may use other means to separate specifiers. In the example in Figure 1, it is possible to identify several specifiers: *"old"*, *"outdoor photography"*, *"distant mountains"*, *"animal documentary"*, *"cinematic"*, *"movie by Dennis Villeneuve"*.
(2) We acknowledge prompt crafting, or prompt engineering, as an **incremental and dynamic process** since its first outputs rarely meet users' expectations and can trigger new ideas along the way. Users usually iterate on prompt specifiers by trial and error, converging or diverging toward their evolving goals.
(3) We assume **the prompt crafting process independent of the generative ML model used**. This assumption is a simplification, as different models may give more or less weight to specific prompt specifiers. Additionally, users have reported difficulties with adjusting their prompting practice to the updated prompts for the newer version of the Midjourney model [53]. In this work, we only analyze addressed to the Stable Diffusion V1 model, but we do not know which model our questionnaire's respondents generally use.
(4) We acknowledge prompting as a **social process**. Midjourney users post their requests on public Discord channels and remain in plain sight of everyone. Some prompt specifiers became viral [27], such as *"trending on artstation"* or *"by Greg Rutkowski"*, despite their minor apparition on training datasets such as LAION-5B [16].

Having established the definitions and assumptions related to TTI generation, the next section presents a study questionnaire designed to examine the community of practice around TTI generation.

## 4 STUDY 1: ONLINE QUESTIONNAIRE

Although previous research offers insights and perspectives on understanding and supporting prompting practices, it does not probe real users' experiences with text-to-image generation. In order to gain a deeper understanding of the sociological traits, usage motivations, and challenges faced by TTI generation practitioners, we conducted a questionnaire study distributed through relevant online communities on social media (Reddit, Discord, and Facebook). This section presents the methods, participants, data collection, and analysis of this questionnaire study and discusses its limitations.

### 4.1 Method and data collection

The online questionnaire broadcasted among AI art online communities focuses on four main aspects (asked in this order):

- **Motivations** for using text-to-image generation;
- **Regularity** of their practice;
- **Usage patterns** through the *complementarity* of text-to-image generation with other tools and techniques, the *challenges encountered* in practice, recurring themes, prompt specifier discovered online, and prompts leading to successful results.
- **Socio-demographic factors**, based on age, gender, and profession in this research. Socio-demographic factors refer to social and demographic attributes that help classify individual respondents within our study.

The questionnaire comprises 10 questions listed in Appendix A.1. Participants were informed that none of the questions were compulsory to complete the questionnaire, as participants might

not be able to share demographic or tips about their practices. Hence, the number of answers varies among questions. We report the type of question and the number of respondents in Table 4 in Appendix A.3. If millions of users use TTI generators, only a fraction congregate in dedicated groups to discuss their practices. For this reason, we published the questionnaire on various AI art groups among three social networks: 10 groups on Facebook, 6 communities on Reddit, and 2 servers on Discord. The questionnaire was shared on Discord's forum channels instead of chat channels, as this approach helped prevent the questionnaire from being overlooked in the presence of numerous messages (including TTI requests) found in chat conversations. The groups on which the questionnaire was disseminated are listed in Table A.2. The responses were gathered between September 25, 2022, and November 27, 2022.

### 4.2 Participants

In this study, 64 individuals (51 men, 7 women, 2 non-binary, and 4 not specified) participated in the online questionnaire. Most participants were committed to text-to-image users, as 47 reported using text-to-image generators almost every day, 15 reported using them several times a week, and 2 reported using them several times a month. The age distribution of the sample is as follows: 27 participants were 45 years of age or older, 12 were between the ages of 35 and 44, 11 were between 25 and 34, 8 were between 18 and 24, and 3 were under the age of 18. The participants represented a diverse range of professions we categorized into 14 sectors. The most frequently represented sectors were Information Technology, with 24 participants, and Art and Culture, with 11 participants. The sample also included non-workers, represented by 7 participants who were students and 5 who were retired. Demographic information of the participants is presented in Figure 2 for better legibility.

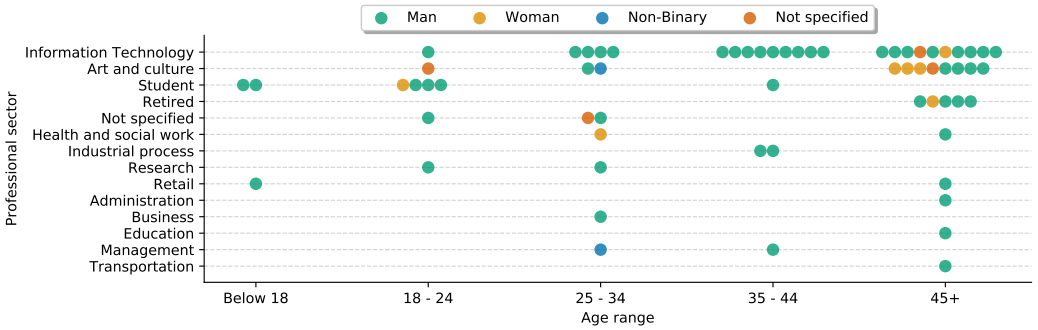



**Fig. 2. Demographic distribution of the 64 respondents. The x axis is the age range, the professional sector (y axis), and gender (color). Each round marker represent a questionnaire respondent.**

### 4.3 Data analysis

The questionnaire comprises 7 open-ended questions. Two of them ask about gender and profession (Question 8 and 9). Responses about gender had low variability and were straightforward to categorize. For professions, we used a simple bottom-up approach to induce broader domains of professional activities and group respondents accordingly.

Questions 1 and 4 are also open-ended questions and deal with **motivations** and **problems** related to the TTI practice. Their responses comprise more variable and long answers. The methodology employed is a simplified thematic analysis approach [17]: all answers were read by the author

to get familiar with the data. Themes were then inductively identified from the first 32 participants, and the response of the remaining 32 participants was then examined according to those themes. Note that participant responses often fall into several of these themes as reported in Figures 12a and 12c in Appendix B. Themes and responses were reviewed and occasionally merged to form the categories presented as bar charts in Figure 3 and 5. These themes are presented and detailed with regard to the research question in Section 4.4.

Additionally, question 3 was a multiple choice question with an open-ended option "other". 11 participants filled the "other" option. Among them, one wrote an existing proposition, 7 proposed new tools or techniques (grouped as "Physical techniques', "3D rendering", "Prompt generator using LLM", "Training data management and augmentation"), and 3 wrote vague or unclear responses we could not categorize. Consequently, both predefined choices and new methods are depicted in Figure 9.

We excluded questions 5, 6, and 10 on **usage patterns** from our analysis for two reasons: first, the low number of responses in Question 10 due to a reluctance to share prompts, and second, the large number of prompts were published in the meantime. We thus decided to shift our focus to analyze these large collections of user prompts such as DiffusionDB [57] that was published on the 27th of October.

### 4.4 Results

This section provides the findings from the online questionnaire whose method is introduced in section 4. The structure represents its main findings: TTI generation mainly seems to be a recreational activity from users with narrow socio-demographic characteristics, its usage extends across platforms and beyond request-response interactions, and inherent model limitations and prompt formulation are the main obstacles encountered by its practitioners.

#### *4.4.1 TTI generation seems to be a recreational activity from narrow socio-demographic characteristics.*

*Socio-demographic distribution of respondents.* The questionnaire responses show an unbalanced demographic and professional distribution of respondents, as most respondents identify as men (51/64). Two professional groups stand out: Information Technology (IT), with 24 practitioners out of 64, and Art and culture, with 11 practitioners out of 64. This disparity in professions can be explained by the fact that text-to-image generation is more likely to reach and interest professionals from its two most impacted domains: technology and art. The gender gap among respondents may be due to the under-representation of woman and other gender minorities in IT fields, as the literature report that woman are four times less likely to be IT practitioners [38]. The questionnaire results also align with this trend, as 22 out of 24 IT practitioners identify as men, while there is no clear gender gap in art and culture respondents. The 22 men working in IT hence represent one-third of our respondents. Additionally, the gender imbalance in our questionnaire might be a selection bias reflecting gender distribution on certain social media, as other works investigating computer-mediated discourse on Reddit also reported gender imbalance among respondents [22].

*Motivations for using text-to-image generators.* From the thematic analysis of responses given in question 1, we identified five main **motivations** for using text-to-image generators, whose occurrences are depicted in Figure 3.

**Leisure** is mentioned by 33 participants (over 64) under various formulations: *"hobby"* (P26, P53, P59), *"fun"* (P10, P38, P42),*"the thrill of creating something appealing"* (P52), or *"I just like creating cool art in my free time"* (P55). Interestingly, two participants mentioned the addictive nature of the

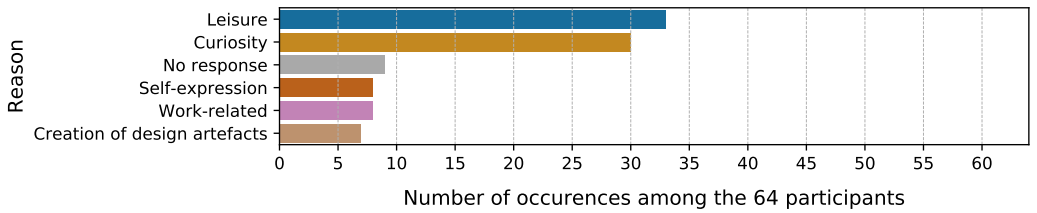



**Fig. 3. Motivations for using text-to-image generation extracted from question 3 of the online questionnaire. The x-axis represents the number of occurrences among the participants' responses as one participant can express several motivations. A participant-level visualization is depicted in Figure 12a, visualized according to age ranges and professional groups.**

practice: *"It is also visually addicting.. and I am OCD can't stop spinning the wheel to see what I get!!"* (P51)

**Curiosity** is mentioned by 30 participants in their response responses. It gathers answers suggesting the desire to learn and discover something new. For instance, participants expressed the desire to comprehend AI abilities that participant 47 best summarizes: *"To learn about the possibilities and limits of AI image generation"* (P47). Two participants mentioned that understanding AI would allow them to anticipate its impact on their professional field (P50) or take advantage of it (P41). For instance, participant 50 said: *"As a game developer, I'm interested how AI could change the game development landscape"* (P50). The second type of curiosity is artistic, as three participants refer to the desire to learn an *"additional tool in an artists toolbox"* (P61) or *"looking for new methods of creating art"* (P22).

**Self-expression** was mentioned by 8 participants as the reason for using TTI generators. These participants view their practice as a valuable form of expression that goes beyond leisure and curiosity. Some common respondents' verbalization include *"self-expression"* (P18, P26) *"delight, artistic expression"* (P4), *"I like to express myself artistically"* (P57), but also more radical formulations: *"to give life to all of the fun and zany ideas trapped inside my head"* (P45), *"almost like a therapy"* (P15).

**Work-related** usages are mentioned by 8 participants, including 3 participants working in the art and culture fields. A painter and digital artist declared that *"this surpasses my skill and gives me a better starting point to assist my own art"* (P51), and a graphic designer for communication stated to *"have occasionally used it as part of my work (which involved making changes to stock photos)"* (P53).

**Creation of design artifacts** is mentioned by 7 participants and comprises usages oriented toward the creation of assets for creative projects, either recreational or professional, such as t-shirt design (P23), music cover (P37), assets for role-play games (P12), prints as gift (P13), and video game assets and textures (P58, P9, P7).

Overall, our findings suggest that TTI generation is primarily a recreational pursuit of users from narrow socio-demographic groups, mostly adult men working in IT and practitioners in the art and culture industries. Text-to-image generators are also used for producing design artifacts for work-related or personal projects, although less frequently. Interestingly, a few participants reported using TTI for extreme reasons, described as an addictive or quasi-therapeutic practice.

*4.4.2 TTI generation practices extend across platforms and beyond request-response interactions.* TTI generator usage can be comprehended when considering its practice within the context of the continuously evolving ecosystem of tools and techniques for TTI developed by researchers and

hobbyists. For this reason, we report the results of question 3 from the questionnaire, which asked about using auxiliary tools for TTI. The techniques reported comprise the pre-selected options from the multiple choice question and additions from participants as explained in section 4.3. Figure 4 shows the frequency of technique usage among survey responses.

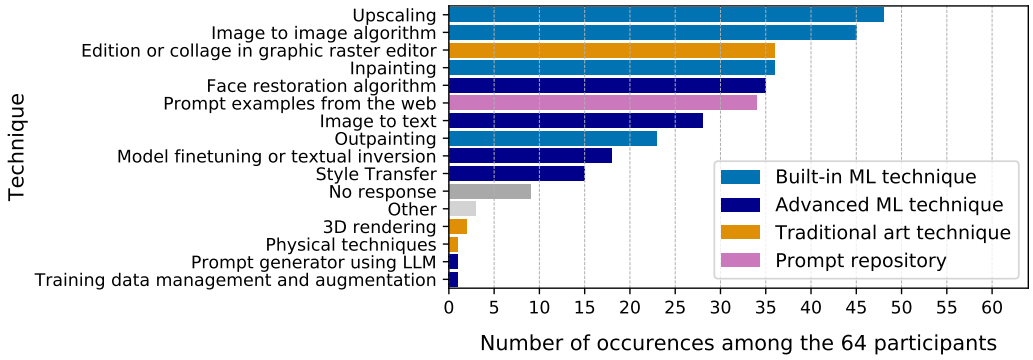



**Fig. 4. Frequency of techniques complementing text-to-image generation among respondents, extracted from question 3 of the online questionnaire. The x-axis represents the number of occurrences among the participants' responses as one participant can choose and provide several techniques. A participant-level visualization is depicted in Figure 12b, visualized according to age ranges and professional groups.**

Four groups of techniques were identified:

**Built-in ML techniques** encompass all tools and techniques that utilize models to manipulate the image or prompt and are commonly found within popular TTI platforms like Midjourney, DALL-E, and DreamStudio (GUI of Stable Diffusion). These techniques include upscaling, image-to-image generation, inpainting, and outpainting. More than half of the participants reported using the first three techniques in their practice.

**Advanced ML Techniques** are ML methods not typically found in standard TTI platforms but rather on public coding repositories such as Google Colab, HuggingFace, and Github. These techniques include face restoration algorithms (e.g. GFPGAN), model fine-tuning and textual inversion (e.g. Dreambooth), style transfer, prompt generation using large language models, and data management and augmentation for training. Although being expected to be used by IT professionals, these advanced ML techniques are also utilized by individuals from diverse socio-demographic attributes, as shown in Figure 12b in Appendix B.

**Traditional art techniques** encompass techniques and tools from art forms that pre-date AI art, such as digital painting, image post-processing, and physical forms of art, such as *"traditional painting using AI generated art as a reference"* (P14). Graphic raster editor software (with Gimp, Photoshop, Procreate) is the most widely used conventional technique from this category. Participants use them not only to edit generated images but also to paint starting images (P37), as images can also be used as input for most TTI models. Two participants, an artist and an IT practitioner, reported using 3D rendering software where generated images are used as textures or objects (P25, P46).

**Prompt platforms** include websites that collect TTI prompts, on which users can reuse and draw inspiration from existing prompts.

Our findings suggest that the TTI generation practice extends beyond just utilizing TTI platforms and request-response interactions. Instead, it involves a range of ML techniques and traditional

tools to complement image generation. Interestingly, advanced ML techniques are not exclusive to IT practitioners but are common tools across users from diverse socio-demographic characteristics.

*4.4.3 Inherent model limitations and prompt formulation are the main obstacles for text-to-image generation.* Understanding the usages of TTI generation in its current state requires examining the obstacles practitioners face. In this section, we present the results of question 4 from our questionnaire, analyzed using the thematic analysis outlined in section 4.3. The analysis identified 8 distinct obstacle categories we describe below, whose frequency among participants is shown in Figure 5.

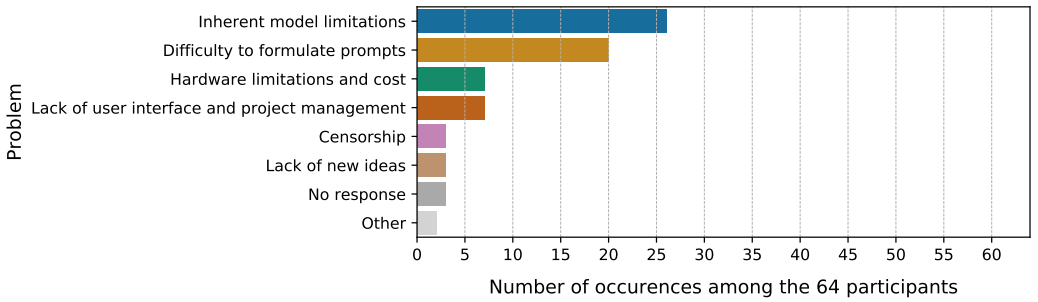



**Fig. 5. Problems encountered for using text-to-image generation, extracted from question 4 of the online questionnaire. The x-axis represents the number of occurrences among the participants' responses as one participant can express several problems. A participant-level visualization is depicted in Figure 12c, visualized according to age ranges and professional groups.**

**Inherent model limitations** is the main issue reported by 25 out of 64 participants. The problems described in this category may stem from model checkpoints i.e., a certain model architecture trained on a specific dataset. The most prevalent issue is the creation of unrealistic artifacts, referred to as *"artifacting"* by participant 53, and mainly affects the generation of hands and facial features. For instance, participant 37 stated that the *"biggest problems are hands. I've had many good images ruined by unusable hands"*. Artifacts can also be watermarks, as mentioned by participant 52, as a result of the presence of stock photo images in the training set.

The second most frequently mentioned inherent model limitation is the model's inability to differentiate between different prompt specifications in the generated image, which we named "concept entanglement". This entanglement was described by 5 participants in various ways. For example, participant 58 reported that *"concepts "leaks", like colors of objects affecting whole image"*, and participant 18 gave a specific example of difficulty when generating an image of *"a devil holding a kitten"* as *"it would always either make the devil a cat, or make the kitten devil like"*.

The third most cited type of inherent model limitation is the model's inability to comprehend niche concepts, mentioned by 4 participants. This problem is related to the text encoder component as *"models are unaware of some concepts, language is restricted to english"* (P58), and *"colloquial terms are not being understood"* (P19). Two techniques listed in question 3 were suggested as a way to address these limitations. Participant 13 mentioned *"training the AI to understand new, niche or highly specific things (like my friends and family) is currently not possible for me."*, while participant 12 claimed that it is *"hard to get specific results, if it is uncommon or new imaginary concept. Photoshop+IMG2IMG solves that"*.

**The difficulty to formulate prompts** is the second most frequently cited obstacle among participants, as expressed by 20 out of 64 participants. Some participants described this obstacle

as independent from the system, with participant 6 stating: *"not being able to properly describe what I have in my mind with words. Language can be a huge limitation."* Other participants viewed this difficulty as a communication issue, highlighting the difficulty in adapting the prompt to the model's "understanding". For example, participant 56 describes prompt formulation as a skill to be learned: *"learning how to talk to the AI so it understands me was a learning curve - I've pretty well got that down now."*

Some participants referred to concept entanglement as a prompt engineering problem. For example, participant 64 stated that *"it's difficult to tell which prompt modifier has the most impact, or how to avoid unintended associations with some keywords"*. Other verbalizations say that *"there are so many adjusting screws that affect each other"* (P28), or insist on the difficulty *"to manipulate the AI into refining its idea into mine"* (P45).

**Hardware limitations and cost** were mentioned by 7 participants. 6 participants expressed frustration with not having a GPU or a powerful enough GPU to run TTI models. One participant mentioned the cost without specifying if it referred to hardware or commercial platform fees.

The **lack of user interface and project management** was identified as an issue by 7 participants. They regret the *"lack of coherence amongst the different tools being built"* (P30) and *"demand for tools with more intuitive user experience"* (P20). Participant 42 stated: *"I want to use the open source tools but they are too hard and there's a LOT of drama"*[2]. Participant 13 mentioned the problem of *"efficiently organising and searching the vast amount of images saved"*.

**Censorship** is a common feature on TTI platforms, as they enforce guardrails to prevent abuse. Guardrails are generally set at both the prompt level by prohibiting certain words and at the generation level by identifying problematic content. Three participants expressed dissatisfaction with the censorship measures. Participant 29 was unhappy with the restriction on NSFW (Not Safe For Work) images, mainly referring to erotic and pornographic content. Participant 51, a painter and digital artist, expressed frustration against the *"parental guidance"* restrictions that prevent the exploration of nudity, a common subject of study in fine arts. *"As a graduate in fine arts we did not have such controls over visual parental guidance as I am seeing in AI now..can not achieve fine art with such constraints"* (P51).

The **other** category comprises two responses that did not fit into the other themes. Participant 17 mentioned the time-consuming aspect of TTI practice, saying *"The only problem I really have is the 24 hours per each day limit."* Meanwhile, Participant 48 expressed concern over the controversy surrounding the status of TTI generation as a new form of art, saying *"the only problem I have is critics thinking that the work is easy."*.

In summary, the main challenge in the TTI generation practice is the inherent model limitations, such as undesirable artifacts and concepts entanglement. Formulating effective prompts is the second main challenge, which can be addressed through learning specialized jargon and a subtle adaptation to the model's outcomes, or in other words, learning to adjust *"screws that affect each other"* (P28). Participants also highlighted the need for better user interfaces and project management tools, providing direct implications for interaction design.

*4.4.4 Summary and limitations.* The online questionnaire study helped to progress toward three research questions. On **RQ1**, the study revealed that TTI generation seems to be a recreational conducted by practitioners from narrow socio-demographic backgrounds. On **RQ2**, the study identified the main challenge associated with TTI generators. On **RQ3**, TTI generation usages extend beyond platforms and request-response interactions, encompassing a range of ML techniques and

[2]The participant likely references the controversy surrounding the developer of an open-source web GUI for Stable Diffusion (AUTOMATIC1111), who was falsely accused of stealing copyrighted code from a text-to-image provider company. The controversy is documented in a Reddit post.

traditional tools. Nevertheless, the online questionnaire method and results also present limitations that are acknowledged in this section.

First, the online questionnaire method is subject to selection and self-selection bias. We gathered responses from various social media platforms to mitigate such biases. Second, online questionnaire generally suffers from sparse or inaccurate responses. If our questionnaire dissemination yielded low response rate, the few respondents had a high level of engagement as they were part of dedicated hobbyist groups in the first place, responded to most of the questions, and often provided extensive responses. Finally, our limited number of respondents (64) may impact the generalizability of our results, particularly for variable data such as socio-demographic factors.

The questionnaire comprised design choices with pros and cons. For instance, question 1 presented examples in parentheses "(leisure, work, curiosity, etc...)" that incentivized participants to give brief answers. While this facilitated the analysis process, it likely resulted in the loss of detail and insights in participants' responses.

The online questionnaire informed on the context of the use of TTI generators but offered limited insights on prompt-crafting patterns and structures. However, we can better understand the prompt crafting process by analyzing the millions of interaction traces (i.e., TTI prompts) generated by this community of practice. Despite the challenges of analyzing variable and unstructured data, examining TTI prompts can counterbalance the limitations of the online questionnaire (i.e., small sample sizes). The following section outlines this analysis and progresses toward establishing a reproducible pipeline for assessing large-scale and collective interactions with generative models.

## 5 STUDY 2: SEMANTIC ANALYSIS OF USER PROMPTS

The questionnaire study provided insight into the socio-demographic characteristics (RQ1), usage challenges (RQ2), and usage patterns (RQ3) of TTI practitioners. This section proposes a closer examination of TTI prompts from real users to further progress toward the third research question on understanding **usage patterns**, in particular on the semantic structure within prompts. To do so, we propose to explore the semantic content of 1.8 million prompts from the DiffusionDB dataset [57] to establish a prompt specifier taxonomy and train a ML model to apply this taxonomy on unseen prompts. This section describes both the creation of a prompt specifier taxonomy and the training of the ML model used to conduct the analysis.

### 5.1 Methods: Taxonomy of text-to-image prompt specifiers

Developing a suitable taxonomy of prompt specifiers is the first step for gaining a higher-level understanding of how prompt are constructed within the community of practice and unravel prompt engineering patterns. The prompt specifier taxonomy proposed in this section was developed from a topic modeling on prompt specifiers, using pretrained neural embeddings for language comprehension. The following sub-sections provide an overview of the data collection and analysis that informed its creation.

*5.1.1 Data collection.* The topic modeling was performed using 1.8 million unique user prompts from the DiffusionDB dataset [57]. As raw prompts were challenging to work with, we focused on individual prompt specifiers instead. For the topic analysis only, prompts were split using commas as separators, as it is the most common practice among practitioners. Only specifiers used at least 100 times and with a length of no more than 35 characters were included, resulting in a set of approximately 7000 specifiers. Table 1 shows the most frequently used specifiers and their frequency among user prompts.

| Rank | Prompt specifier | Occ. (in %) | Rank | Prompt specifier | Occ. (in %) |
|---|---|---|---|---|---|
| 1 | highly detailed | 2,610 | 16 | cinematic lighting | 0,713 |
| 2 | artstation | 2,334 | 17 | 4 k | 0,711 |
| 3 | sharp focus | 2,048 | 18 | 8k | 0,699 |
| 4 | concept art | 1,916 | 19 | detailed | 0,666 |
| 5 | trending on artstation | 1,867 | 20 | photorealistic | 0,599 |
| 6 | intricate | 1,465 | 21 | unreal engine | 0,594 |
| 7 | digital painting | 1,440 | 22 | masterpiece | 0,578 |
| 8 | 8 k | 1,416 | 23 | greg rutkowski | 0,534 |
| 9 | octane render | 1,375 | 24 | 4k | 0,529 |
| 10 | illustration | 1,359 | 25 | realistic | 0,500 |
| 11 | smooth | 1,136 | 26 | artgerm | 0,496 |
| 12 | elegant | 1,071 | 27 | hd | 0,452 |
| 13 | cinematic | 0,918 | 29 | dramatic lighting | 0,446 |
| 14 | digital art | 0,868 | 29 | volumetric lighting | 0,438 |
| 15 | fantasy | 0,743 | 30 | cgsociety | 0,411 |

**Table 1. Ranking of the popular prompt specifiers and their occurrences among 1.8 million unique prompts from the DiffusionDB dataset [57]**

*5.1.2 Data analysis: topic modelling using language models.* The topic modeling analysis follows the approach described by Grootendorst [26]. First, all specifiers were embedded in machine-learned vector representations. We explored two types of text encoders to perform the topic modeling:

(1) The CLIP vision-supervised text encoder, specifically the `ViT-L-14` checkpoint trained by OpenAI and used to train the Stable Diffusion model;
(2) The MPnet embedding (Masked and Permuted Pre-training for Language Understanding), a model designed for language understanding tasks such as information retrieval, clustering, and sentence similarity. We specifically used the `all-mpnet-base-v2` checkpoint trained by Microsoft, which obtained the best average performance for sentence embeddings (evaluated on 14 datasets) and semantic search (evaluated on 6 datasets) [15].

The encoded representations were further reduced in dimension using the U-MAP algorithm [36] to alleviates the curse of dimensionality during the topic analysis. This resulted in projecting the prompt specifiers in a 5-dimensional space to conduct the analysis while avoiding significant information loss due to feature collapse. A hierarchical density-based spatial clustering model (HDBSCAN) was used to cluster the data points i.e., prompt specifiers. Unlike other clustering methods, HDBSCAN does not require specifying the number of clusters, can handle non-linear distributions, and identify outliers as a separate category. Finally, we ran a class-specific term frequency–inverse document frequency (c-TF-IDF) analysis [44] to select the top candidates that specifically describe a cluster. Cluster titles were manually chosen based on those candidates and combined if considered too similar.

We visualized the two topic models (CLIP and MPnet) in interactive 2D maps (further reduced with U-MAP), depicted on Figure 6[3]. The maps display all prompt specifiers as points with labels when hovered over and cluster titles placed at the centroid of each cluster. Users can navigate the map by zooming in or out.

[3]The interactive map is available at https://teo-sanchez.github.io/demos/prompting_map.

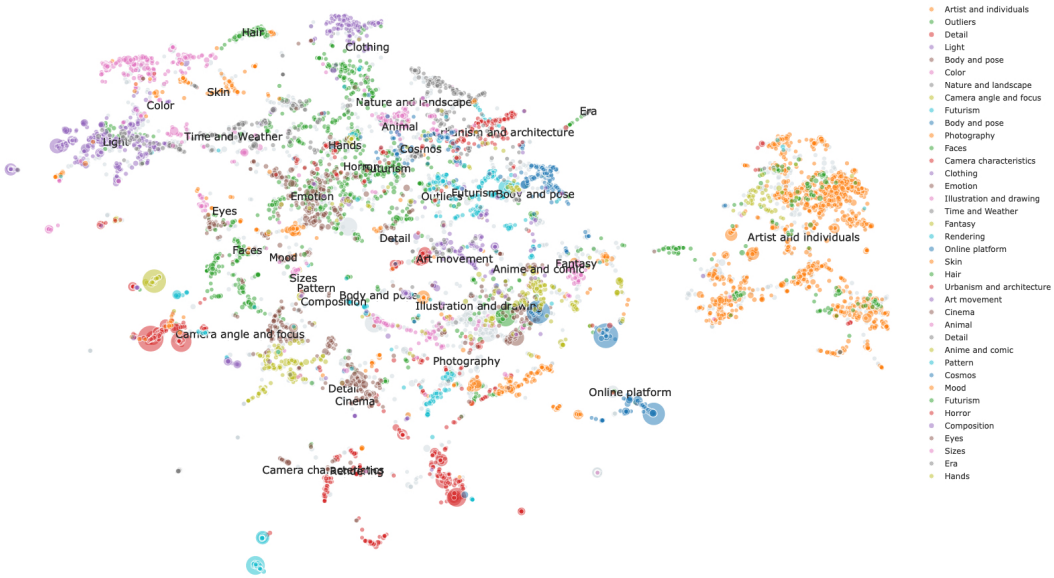



**Fig. 6. Interactive visualization of prompt specifiers in a 2-dimensional reduction of the `all-mpnet-base-v2` embedding. The visualization also represents the 38 topics automatically extracted from the neural topic modeling approach [26].**

A salient outcome was the difference in clarity and organizational coherence between the CLIP and MPNet visualizations. The CLIP representations and clusters are more intertwined and challenging to comprehend than MPNet. We hence focused our analysis solely on MPNet topic modeling. The automatic clustering using MPNet provided 38 topics and a ranking of the most representative prompt specifiers for each category. The main author renamed each topic with a single title. The map and categories were shared with colleagues which collectively discussed a suitable taxonomy using an inductive approach. Categories that were too specific or relative to a subject were discarded (e.g. Hair, Eyes, Cosmos etc.) while the remaining categories were grouped to form broader topics. For instance, *Camera angle and focus*, *Photography*, *Futurism*, *Anime and comics*, *Fantasy*, and *Art movement* were grouped as *Genres*. The final taxonomy was hence reduced to 16 sub-categories, including *Subject*. Categories describing common concepts were grouped to describe higher-level semantics of TTI prompts: *Photography and cinema*, *Painting*, *Rendering*, and *Illustration* were grouped into *Mediums*. Similarly, *Artists*, *Art genres and movements*, *Artwork*, and *Art repositories* were grouped as *Influences*, and *Era*, *Weather*, and *Emotions* were grouped as *Context*.

*5.1.3 Prompt specifier taxonomy.* This subsection presents the taxonomy obtained from the topic analysis of prompt specifiers in the DiffusionDB dataset. The taxonomy is shown in Figure 7 and consists of 8 main categories:

**Subject** refers to elements forming the primary focus of the prompt. Subjects are not always specified in real prompts, and we decided to consider them in our taxonomy as it is deemed important for the construction of assistive tools for prompting. In the example in Figure 1, the subject is *"An old groundhog wearing a cloak and an ivy cap"*.

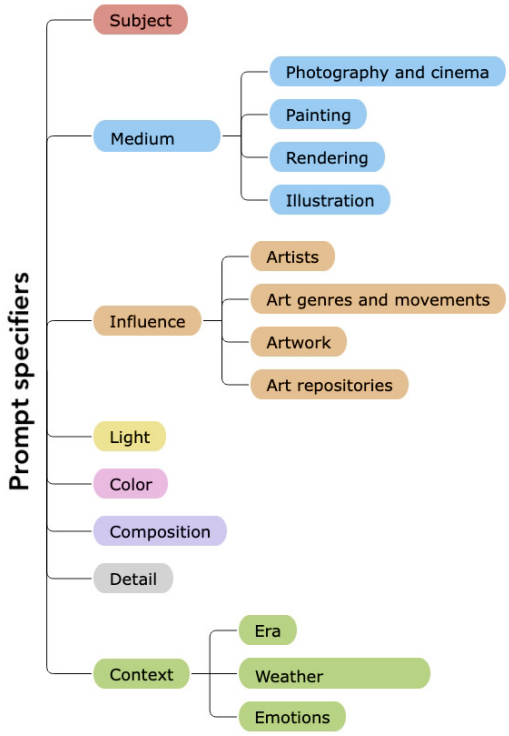



**Fig. 7. Prompt specifier taxonomy developed inductively from a topic modeling of user prompts with ML language models (`all_mpnet_base_v2`).**

**Medium** encompasses specifiers evoking the artistic media found in TTI prompts: photography and cinema, painting, rendering, and illustration (that includes digital art). From the example shown in Figure 1, *"outdoor photography"*, *"cinematic"*, and *"movie"* belong to this category.

**Influence** encompasses external and artistic influences referenced in the prompts. It includes four subcategories: artist names, art genres and movements, artwork names, and art repositories (primarily online platforms for digital artists). The last subcategory is used to improve the quality and composition of the generated image. For instance, specifiers such as "trending on artstation" are expected to force the model to access the learned information on the best visual art encountered on Artstation. This practice can be qualified as "quality boosters" according to the taxonomy from [39].

**Light** encompasses specifiers that describe the lighting conditions in the scene. The most common ones are *"dramatic lighting"*, *"studio lighting"*, and *"emotional light"*.

**Color** encompasses specifiers that describe the colors of objects or global color schemes. The most common ones are *"colorful"*, *"vibrant colors"*, and *"vivid colors"*.

**Composition** refers to specifiers describing the overall arrangement of elements in the image. The most common ones are *"epic composition"*, *"cinematic composition"*, and *"beautiful composition"*.

**Detail** encompasses specifiers aimed at enhancing the details of the generated image. The most popular specifiers from this category are *"highly detailed"*, *"sharp focus"*, and *"intricate"*. They also belong to "quality boosters" according to the taxonomy outlined in [39].

**Context** encompasses specifiers that provide contextual information that cannot be fully represented, such as the era (*"in the 1920s"*), the weather (*"fog"*), and emotions (*"epic"*).

*5.1.4 Limitations.* The taxonomy presented in section 7 was developed through a process that integrated automatic topic modeling with human-made design decisions. The methodology used in the construction of the taxonomy involved the merging of categories, naming of groups, and restructuring of the overall classification system. The taxonomy does not pretend to depict all specifiers, and one can easily find examples that challenge the choices made. For example, *"renaissance"* may refer to an era but is more likely to describe the corresponding art movement in the context of TTI generation. Our contribution is to provide a practical taxonomy that can be used to analyze the semantic content of TTI prompts further and derive prompting support tools. If the present taxonomy was constructed from the analysis of text prompts, alternative approaches could draw from the field of art history and iconography to describe the semantic content of TTI prompts. For instance, Burford et al. [18] proposed a user-centered taxonomy based on how people perceive and interpret images, from direct sensory elements to high-level abstractions.

### 5.2 Methods: Prompt specifier entity recognition model

The previous sub-section presented the construction of a practical taxonomy for text-to-image prompt specifiers. The second step in the semantic analysis of TTI prompts involves automating the taxonomy to a large collection of prompts in order to gain an understanding of semantic patterns in the prompt construction and provides insights into user behavior. For this purpose, we describe the training of `en_ner_prompting`, a machine learning (ML) model that can parse prompts and identify specifiers according to the classes from the taxonomy. The model is available under the BY 3.0 Creative Common license at the following address: https://huggingface.co/teo-sanchez/en_ner_prompting. Overall, this publicly-available model contributes to establishing a reproducible pipeline for analyzing large corpora of TTI prompts. The training pipeline and all configuration files are available at https://github.com/teo-sanchez/ner_prompting. The following sections describe the training data collection and the chosen ML pipeline.

*5.2.1 Training data.* The training and evaluation set was created using annotated text prompts from DiffusionDB [57]. The author of the paper carried out the annotation process using an annotation tool called Prodigy [24], which uses active learning techniques to sample examples to annotate optimally. The annotation was performed in two phases. In both phases, the annotation interface displays the text prompts and the corresponding images generated by users with the Stable Diffusion model. In the first phase, text prompts were annotated to train a preliminary Name Entity Recognition (NER) model. In the second phase, the model predictions were displayed and corrected to retrain improved model versions more rapidly. A total of 715 text prompts were annotated using the 16 classes from the taxonomy outlined in section 5.1.3 (counting sub-categories as independent categories). The annotation was stopped as the model performance stagnated with additional annotations.

*5.2.2 Machine Learning pipeline.* The following section describes the machine learning (ML) pipeline used to train `en_ner_prompting`, a model capable of extracting prompt specifiers from the 16 classes from the taxonomy outlined in section 5.1.3 from unseen text-to-image prompts. The model is a pipeline of two components:

**Tok2vec encoder component**: A tok2vec encoder was finetuned using raw prompts from DiffusionDB starting from the pre-trained tok2vec component of `en_core_web_lg` designed by SpaCy [51]. This component is an ML model that learns to produce suitable and dynamic vectors for tokens by analyzing their lexical attributes. The unsupervised finetuning of this tok2vec component

on raw user prompts will likely produce specific and more reliable embedding vectors and improve downstream tasks performances.

**NER component**: The Name Entity Recognition model was trained on top of the tok2vec representations using annotated data and the 16 classes in taxonomy. The NER model chosen relies on a transition-based parsing approach [33], which builds state transitions to condition the predictions. Despite performances probably inferior to modern transformers' neural architectures, this model can run in real-time on a CPU, making it a candidate for building interactive support systems for TTI generation. Therefore, the trained model can serve not only as an analytical tool but also as a technological brick to be used in interactive systems.

The NER training was performed on 80% of the annotated prompts, which account for 572 instances. All training data and configuration files for the tok2vec pretraining and the NER training can be found in the project repository https://github.com/teo-sanchez/ner_prompting.

*5.2.3 Model evaluation.* This subsection reports the `en_ner_prompting`model evaluation. An example of model prediction on a realistic user prompt is shown in Figure 8. It was computed in 0.0003 seconds on CPU, demonstrating that the current model can be used for real-time implementation of interactive systems.

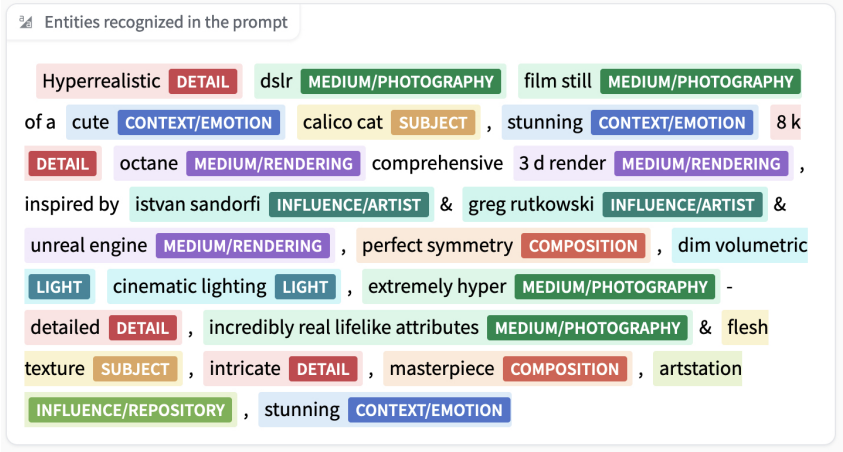



**Fig. 8. Example of prediction from the `en_ner_prompting` model on a real user prompt.**

The model was evaluated on 20% of the data excluded for training (143 prompts) and achieved an F-score of 0.73. The global and class-specific performance metrics can be found in Table 2. It should be noted that the model performance varies across classes. The best-recognized classes are "Detail" (F-score=0.91), "Light" (F-score=0.89), and "Influence/Artists" (F-score=0.88), while the worst are "Context/Weather" (F-score=0.50), "Subject" (F-score=0.55), and "Influence/Artwork" (F-score=0.55). The poor recognition of subjects may be due to their variable length and the possibility of spanning multiple words in the prompt. Artworks are also challenging to recognize as they are rarely mentioned in TTI prompts and can easily be mistaken as subjects.

## 5.3 Results: Semantic analysis of user prompts

In this section, we present an exploratory analysis of DiffusionDB using the `en_ner_prompting`model trained according to the taxonomic framework outlined in section 5.1.3. We ran the model on 1.8 million unique prompts from DiffusionDB and published the processed data at the following link https://huggingface.co/datasets/teo-sanchez/diffusiondb_ner.

| | F-score | Preci. | Recall |
|---|---|---|---|
| **Global performance** | 0.73 | 0.74 | 0.73 |
| subject | 0.55 | 0.56 | 0.54 |
| medium/photography | 0.72 | 0.77 | 0.68 |
| medium/painting | 0.85 | 0.92 | 0.79 |
| medium/rendering | 0.85 | 0.92 | 0.79 |
| medium/illustration | 0.80 | 0.88 | 0.73 |
| influence/artists | 0.88 | 0.86 | 0.90 |
| influence/genres | 0.69 | 0.63 | 0.76 |
| influence/artwork | 0.55 | 0.75 | 0.43 |
| influence/repositories | 0.80 | 0.93 | 0.89 |

| | F-score | Preci. | Recall |
|---|---|---|---|
| light | 0.89 | 0.87 | 0.92 |
| color | 0.76 | 0.75 | 0.77 |
| composition | 0.63 | 0.67 | 0.59 |
| detail | 0.91 | 0.93 | 0.89 |
| context/era | 0.80 | 0.73 | 0.89 |
| context/weather | 0.50 | 0.57 | 0.44 |
| context/emotion | 0.67 | 0.70 | 0.65 |

**Table 2. Global and class-specific performance metrics of the trained `en_ner_prompting` model.**

*5.3.1 Prevalence of prompt specifier's type.* Figure 9 illustrates the prevalence of the prompt specifier's types within the DiffusionDB dataset, as predicted by the `en_ner_prompting` model. The figure represents the frequency of each class of prompt specifier among the 1.8 million unique prompts.

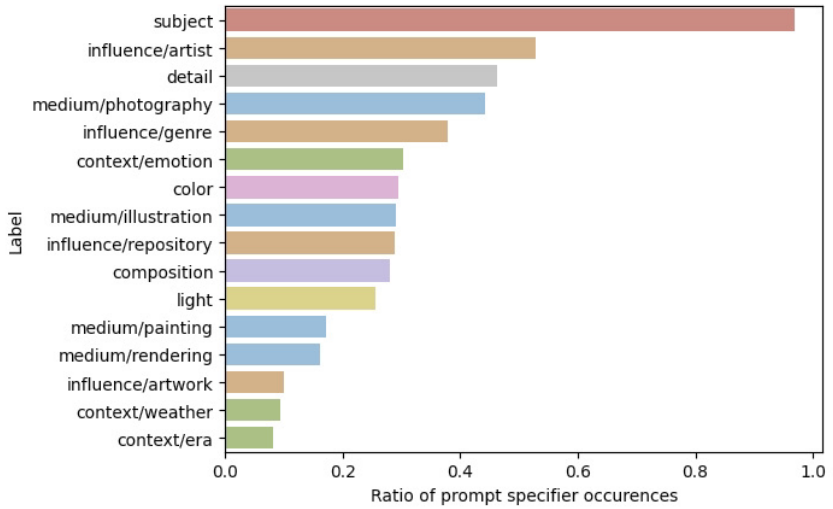



**Fig. 9. Frequency of the type of prompt specifiers among the unique prompts of DiffusionDB [57]**

From the figure, it can be observed that our model detected a subject in the majority of TTI prompts (97%). Additionally, artist names were detected in over half of the user prompts (52%). Lastly, specifiers that provide details, mainly "quality boosters", are the most frequently used (46%) after artist names.

*5.3.2 Transition probability between prompt specifiers.* By applying the `en_ner_prompting`model on DiffusionDB, we counted transitions between each types of specifiers and estimated transition probabilities.

Figure 10 presents a network diagram showing the most probable specifier transitions encountered in user TTI prompts. The figure shows that artist names tend to be mentioned in sequences,

with one artist's name following another in 62% of cases. Similarly, detail specifiers or "quality boosters" are frequently mentioned in groups or sequences, with one detail specifier following another in 26% of cases. The other transitions indicate that, just before the subject, descriptions primarily focus on attributes such as color, era, emotions, and genre.

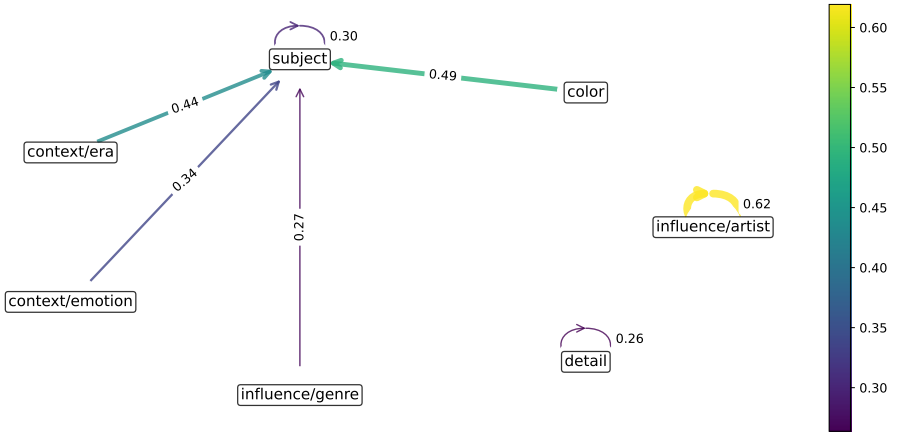



**Fig. 10. The 7th most probable transitions between prompt specifiers' types.**

### 5.4 Summary and limitations

This section presented a method to analyze the semantics of users' prompts. To do so, we established a prompt specifier taxonomy and trained `en_ner_prompting`, a lightweight model for recognizing prompt specifiers based on this taxonomy. The model can uncover usage patterns, specifically the semantic patterns in how TTI practitioners craft their prompts. The results reveal that artist names are the most common specifiers and are often used in a sequential manner.

In addition, the model demonstrates efficient and relatively accurate recognition of prompt specifiers, with only a few hundred training examples. The model is particularly accurate at identifying artist names. More importantly, the model is usable in real-time to build interactive systems assisting prompting practices.

Although analyzing data using predictions from an imperfect model may introduce some degree of uncertainty, this analysis still holds value by offering novel quantitative insights into the practice of TTI prompting. These findings complement and enrich those from the first study, thereby providing a more comprehensive understanding of the subject matter.

## 6 DISCUSSION

This section discusses the results from both the questionnaire and the prompt specifier analysis, emphasizing how our findings can assist in developing tools that support the discovery of new artistic jargon and democratize artistic self-expression. Additionally, we discuss the ethical concerns surrounding the recognition of human labor involved in training TTI generators, and our contributions offer insights into building systems to identify such labor.

### 6.1 Implications for designing interactive assistant for prompt crafting

The questionnaire's results indicate that the main obstacles for TTI practitioners are inherent model limitations and difficulties in formulating prompts. While the first obstacle is more a matter of

ML than HCI research, the second offers promising opportunities for HCI research, which will be discussed in this section. The `en_ner_prompting`developed in this research, intended to analyze large prompt corpora, also responds to the design guidelines of Dang et al. [20] as it can be used to parse and extract prompt specifiers. An interactive system could either suggest alternatives for typed specifiers as illustrated in Figure 11a, or suggest missing categories of prompt as illustrated in Figure 11b. The categories of prompt specifiers may act as a typing system to parse and complement a prompt with new elements.

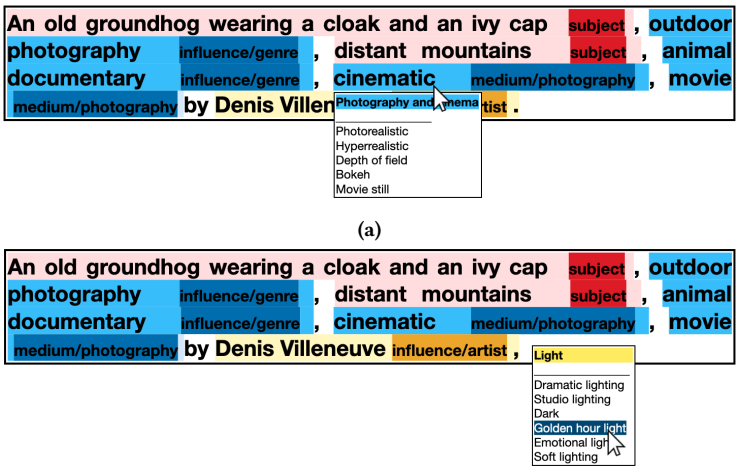



**Fig. 11. Illustrating prompting assistance through the suggestion of prompt specifier alternatives (a) and continuation with new types of specifiers (b).**

The key question is now: Which specifier should be recommended to users? A community-focused approach, such as suggesting the most popular specifiers, might help onboard users unfamiliar with artistic terms and references. Our findings suggest that TTI has a potential for self-expression that needs to be fully understood. Further research is needed to investigate how TTI can democratize new forms of expression and communication, especially among the under-represented socio-demographic groups that might be present in our respondent pool (e.g., elderly, children).

An alternative "community-aware" approach would be to recommend specifiers leading to unexplored regions of the latent space. This type of assistance would benefit users looking to take advantage of the unpredictable nature of generative systems, probably artists and researchers. A last approach would be to design "model-aware" recommendations to address the "entanglement" problems reported by questionnaire participants. The system would then identify specifiers prone to generating confusing or low-aesthetic images and suggest more appropriate ones. Overall, the design of prompt specifier recommender tools should account for the varied motivations and characteristics of TTI practitioners.

### 6.2 Recognizing human artists' work in text-to-image generation

Our results show that artist names are the primary type of specifiers used in TTI prompts. However, the widespread use of text-to-image models sparked controversy among professional artists. They accused companies of profiting from legal ambiguity by using copyrighted work in public training datasets (e.g. LAION-5B [47]), and expressed concern about potential job displacement in the art

industry [55]. In particular, artists voiced their discontent against the online art gallery platform Artstation for allowing AI-generated artwork to be posted on the platform [48]. In early 2023, legal actions were taken against AI art companies by stock image company Getty Images in the UK [32] and independent artists in the US [30]. The outcome of these trials might redefine the rights and protections of AI-generated media.

If most ethical concerns related to TTI are legal matters, technical advances might mitigate some issues. First, artists should have the option to opt out of TTI training data. This approach is being investigated by StabilityAI but will only apply to future models rather than existing ones that are now widely available online. Second, opted-in artists might be compensated when their names are used in user prompts, similar to how musicians are paid for streams on music platforms. This approach could be a business model recognizing human contributions to AI-generated synthetic media. Our work informs and contributes to this perspective as we found that 1) artist names are the main type of prompt specifier used after the main subject, and 2) we provide a ML model that can reliably identify artist names in TTI prompts.

## CONCLUSION

The present work examined the young community of practice that developed around text-to-image (TTI) generation. A questionnaire study was conducted among various online groups on TTI generation. Our findings revealed the motivations, challenges associated with the use of TTI generators, and socio-demographic characteristics of this community. In particular, we discovered that TTI generation seems to be a recreational activity from practitioners with narrow socio-demographic characteristics, mostly male technologists and artists. Additionally, TTI practitioners use various auxiliary techniques across platforms and beyond request-response interactions. Available systems lack coherent user interfaces and project management. Lastly, inherent model limitations and prompt formulations are the main obstacles encountered by TTI generation practitioners.

We conducted an exploratory analysis of 1.8 million TTI prompts from the DiffusionDB dataset [57] to better comprehend usage patterns of TTI generation. We took a topic modeling approach using pre-trained ML models for language understanding. Our results informed the construction of a prompt specifier taxonomy and a corresponding model for extracting semantic categories from TTI prompts in real-time. This model helped us understand the semantic structure of TTI prompts and can alternatively act as a technology brick to design meaningful interactions with TTI generators. We found that artist names are the main type of prompt specifier used and tend to be invoked sequentially.

In conclusion, this work provides a new perspective focused on human practices, which can benefit HCI and ML. Beyond the research perspectives, we hope our work will inform the public debate on TTI generation and progress toward meaningful usage of this technology.

## ACKNOWLEDGMENTS

We would like to thank C&C chairs and anonymous reviewers for their efforts in reviewing this paper and their constructive comments. We gratefully acknowledge the Banque Publique d'Investissement (BPI) France for funding this research through the deep tech development aid funding, particularly Mortimer Pignon for his support in the project. I thank the Selas Studio members for their interest and assistance through the research process: Léonard Strouk, Alexandre Lavallée, Benjamin Trom, Roméo Incardona, and Antoine Aparicio. Their collaboration was instrumental in the success of this research. Thank you to Baptiste Caramiaux for his proofreading efforts and insightful advice during the early stage of the paper.

## REFERENCES


[1] 2012. Common Crawl. Available at https://commoncrawl.org.
[2] 2022. DALL·E Editor Guide | OpenAI Help Center. https://help.openai.com/en/articles/6516417-dall-e-editor-guide
[3] 2022. DALL·E: Introducing Outpainting. https://openai.com/blog/dall-e-introducing-outpainting/
[4] 2022. Imagine Parameters Illustrated - Midjourney Documentation. https://midjourney.gitbook.io/docs/imagine-parameters#image-prompting-with-url
[5] 2022. Lexica. https://lexica.art/
[6] 2022. Midjourney. Available at https://www.midjourney.com.
[7] 2022. Midjourney prompt generator - promptoMANIA. https://promptomania.com/midjourney-prompt-builder/
[8] 2022. MidJourney Prompt Tool. https://prompt.noonshot.com/midjourney
[9] 2022. MidJourney Random Commands Generator. https://blog.user.today/midjourney/
[10] 2022. Phraser — the collaborative creative AI tool. https://phraser.tech/builder
[11] 2022. Prompt Battle. https://promptbattle.com/
[12] 2022. Prompt Hunt. https://www.prompthunt.com/explore
[13] 2022. Prompt Silo. https://pheeds.com/PromptSilo.php?ref=futuretools.io
[14] 2022. PromptBase | Prompt Marketplace: DALL·E, Midjourney, Stable Diffusion & GPT-3. https://promptbase.com/
[15] 2023. Pretrained Models — Sentence-Transformers documentation. https://www.sbert.net/docs/pretrained_models.html?highlight=pretrained
[16] Andy Baio. 2022. Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator - Waxy.org. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/
[17] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, 57–71. https://doi.org/10.1037/13620-004
[18] Bryan Burford, Pam Briggs, and John P. Eakins. 2003. A Taxonomy of the Image: On the Classification of Content for Image Retrieval. *Visual Communication* 2, 2 (2003), 123–161. https://doi.org/10.1177/1470357203002002001
[19] Baptiste Caramiaux and Sarah Fdili Alaoui. 2022. "Explorers of Unknown Planets": Practices and Politics of Artificial Intelligence in Visual Arts. *Proceedings of the ACM on Human-Computer Interaction* 1, 1 (2022), 1–24. https://hal.inria.fr/hal-03762351%0Ahttps://hal.inria.fr/hal-03762351/document
[20] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models; How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creat. (2022). https://doi.org/10.1145/nnnnnnn.nnnnnnn
[21] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2022. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. *arxiv.org* (2022). http://arxiv.org/abs/2212.07476
[22] S Craig Finlay and Franklin D Schurz. 2014. Age and gender in Reddit commenting and success. (2014). https://doi.org/10.1633/JISTaP.2014.2.3.2
[23] Paul Ginsparg. 2011. ArXiv at 20. *nature.com* (2011). https://www.nature.com/articles/476145a
[24] ExplosionAI GmbH. 2019. Computer Vision ·Prodigy ·An annotation tool for AI, Machine Learning & NLP. https://prodi.gy/%0Ahttps://prodi.gy/features/computer-vision
[25] Dejan Grba. 2022. Deep Else: A Critical Framework for AI Art. *mdpi.com* (2022). https://doi.org/10.3390/digital2010001
[26] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (3 2022). http://arxiv.org/abs/2203.05794
[27] Melissa Heikkilä. 2022. This artist is dominating AI-generated art. And he's not happy about it. | MIT Technology Review. https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/
[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 2020-December (6 2020). https://doi.org/10.48550/arxiv.2006.11239
[29] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research* 23 (2022). https://www.jmlr.org/papers/volume23/21-0635/21-0635.pdf
[30] Irina Ivanova. 2023. Artists sue AI company for billions, alleging "parasite" app used their work for free - CBS News. https://www.cbsnews.com/news/ai-stable-diffusion-stability-ai-lawsuit-artists-sue-image-generators/
[31] Anna Kantosalo and Hannu Toivonen. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. *Proceedings of the 7th International Conference on Computational Creativity, ICCC 2016* (2016), 77–84. https://www.computationalcreativity.net/iccc2016/wp-content/uploads/2016/01/Modes-for-Creative-

Human-Computer-Collaboration.pdf
[32] Jennifer Korn. 2023. Getty Images suing the makers of popular AI art tool for allegedly stealing photos | CNN Business. https://edition.cnn.com/2023/01/17/tech/getty-images-stability-ai-lawsuit/index.html
[33] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference* (2016), 260–270. https://doi.org/10.18653/v1/n16-1030
[34] Vivian Liu and Lydia B. Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. *Conference on Human Factors in Computing Systems - Proceedings* 1, 1 (4 2022), 1–27. https://doi.org/10.1145/3491102.3501825
[35] Tiago Martins, João M. Cunha, João Correia, and Penousal Machado. 2023. Towards the Evolution of Prompts with MetaPrompter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13988 LNCS (2023), 180–195. https://doi.org/10.1007/978-3-031-29956-8{_}12
[36] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2 2018). http://arxiv.org/abs/1802.03426
[37] Anna Notaro. 2020. State of the Art: A.I. through the (artificial) artist's eye. BCS Learning and Development Ltd. https://doi.org/10.14236/ewic/EVA2020.58
[38] Organisation for Economic Co-operation OECD. 2018. Bridging the digital gender divide: Include, upskill, innovate. *voced.edu.au* (2018). https://www.voced.edu.au/content/ngv:81069
[39] Jonas Oppenlaender. 2022. A Taxonomy of Prompt Modifiers for Text-To-Image Generation. *arxiv.org* 1, 1 (4 2022), 1–15. http://arxiv.org/abs/2204.13988
[40] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. *25th International Academic Mindtrek conference* (11 2022), 192–202. https://doi.org/10.1145/3569219.3569352
[41] Nikita Pavlichenko and Dmitry Ustalov. 2022. Best Prompts for Text-to-Image Models and How to Find Them. (9 2022). https://doi.org/10.48550/arxiv.2209.11711
[42] @pharmapsychotic. 2022. CLIP Interrogator - a Hugging Face Space by pharma. https://huggingface.co/spaces/pharma/CLIP-Interrogator
[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. (2 2021). https://doi.org/10.48550/arxiv.2103.00020
[44] Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning* 242, 1 (2003), 29–48. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c
[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. (8 2022). https://doi.org/10.48550/arxiv.2208.12242
[46] Gustavo Santana. 2022. MagicPrompt Stable Diffusion - a Hugging Face Space by Gustavosta. https://huggingface.co/spaces/Gustavosta/MagicPrompt-Stable-Diffusion
[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. (10 2022). https://doi.org/10.48550/arxiv.2210.08402
[48] Ali Shutler. 2022. Artists protest use of AI-generated artwork on portfolio site Artstation. https://www.nme.com/news/gaming-news/artists-protest-use-of-ai-generated-artwork-on-portfolio-site-artstation-3366778
[49] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *32nd International Conference on Machine Learning, ICML 2015* 3 (3 2015), 2246–2255. https://doi.org/10.48550/arxiv.1503.03585
[50] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems* 32 (7 2019). https://doi.org/10.48550/arxiv.1907.05600
[51] SpaCy. 2020. spaCy · Industrial-strength Natural Language Processing in Python. https://spacy.io/
[52] Luke Stark and Kate Crawford. 2019. The work of art in the age of artificial intelligence: What artists can teach us about the ethics of data practice. *Surveillance and Society* 17, 3-4 (2019), 442–455. https://doi.org/10.24908/ss.v17i3/4.10821
[53] Xander Streenbrugge. 2022. Xander Steenbrugge on Twitter: "One annoying thing about Generative AI right now is that new models are constantly emerging, and each one requires you to "relearn how to prompt" it. Prompts that looked amazing in SD v1.5 don't in v2.0. This constant relearning feels like a struggle we can prob improve on" / Twitter. https://twitter.com/xsteenbrugge/status/1595780305981689862

[54] Succintly AI. 2022. succinctly/text2image-prompt-generator · Hugging Face. https://huggingface.co/succinctly/text2image-prompt-generator

[55] Chris Vallance. 2022. "Art is dead Dude" - the rise of the AI artists stirs debate - BBC News. https://www.bbc.com/news/technology-62788725

[56] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. *openaccess.thecvf.com* (2021). http://openaccess.thecvf.com/content/CVPR2021/html/Wang_Towards_Real-World_Blind_Face_Restoration_With_Generative_Facial_Prior_CVPR_2021_paper.html

[57] 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. (10 2022). https://doi.org/10.48550/arxiv.2210.14896

# A ONLINE QUESTIONNAIRE

## A.1 List of questions

1. Please describe why do you make AI art? (Leisure, work, curiosity, etc...) (Open-ended answer)

2. How often do you make AI art?
- ◯ Almost every day
- ◯ Several times a week
- ◯ Several times a month
- ◯ Several times a month
- ◯ Less than once a month

3. Do you complement text-to-image practices with other tools? If yes, precise: (NB: options are randomized)
- □ Upscaling algorithms
- □ Style transfer algorothms
- □ Inpainting
- □ Outpainting
- □ Image-to-text algorithms (e.g. clip interrogator)
- □ Model finetuning or textual inversion
- □ Edition or collage in graphic softwares (e.g. Photoshop)
- □ Browse examples from repository (e.g. Lexica.art or Midjourney)
- □ Face restauration algorithm (e.g. GFPGAN)
- □ Other: (Open-ended answer)

4. What problem(s) do you face when creating with generative AIs? Please describe them in your own words:
(Open-ended answer)

5. Do you have recurring themes in your creations? If yes, please write one theme per line.
(Open-ended answer)

6. What are the prompt specifiers you discovered online and now use frequently? (e.g. trending on artstation) Please write one prompt specifier per line. (Open-ended answer)

7. Age?
- ◯ Below 18
- ◯ 25 - 34
- ◯ 35 - 44
- ◯ 45+

8. Gender? (Open-ended answer)

9. Profession or field of activity? (Open-ended answer)

10. Could you share one of the prompt you are proud of? (Open-ended answer)

### A.2 Questionnaire dissemination

The responses were gathered between September 25, 2022, and November 27, 2022.

| Social Network | Name of the group |
|---|---|
| Reddit | r/StableDiffusion |
| | r/MidJourneyDiscussions |
| | r/aiArt |
| | r/midjourney |
| | r/AIArtwork |
| | r/aiARTistsUNITE |
| Facebook | AI Artists |
| | Stable Diffusion |
| | Promptism |
| | AI Generated Art |
| | STABLE DIFFUSION ARTIST COMMUNITY |
| | MidJourney AI Art Station |
| | Prompt whispering |
| | AI Art Universe |
| | Midjourney Official |
| | Promptlib - AI Art Library |
| Discord | Stable Diffusion - prompting-help channel |
| | Midjourney - Community forums |

**Table 3. Social network groups utilized for questionnaire dissemination**

### A.3 Question types and number of respondent

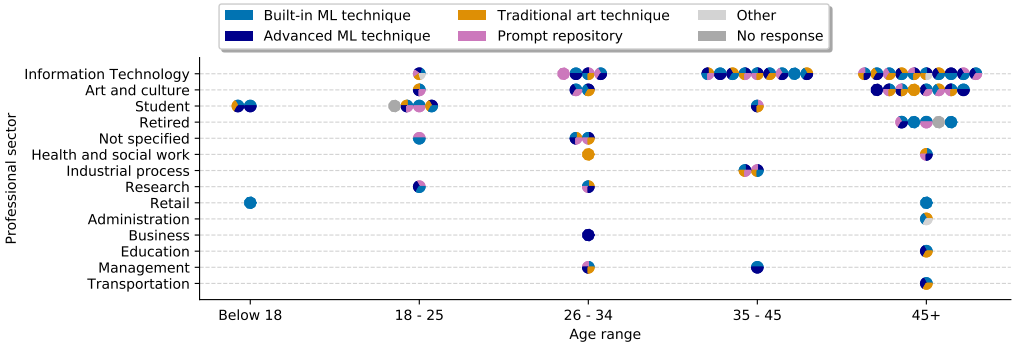
| No. | Question summary | Type | # Respondent |
|---|---|---|---|
| 1 | Why do you make AI art? | Open-ended | 55 |
| 2 | How often? | Multiple choice with exclusive answers | 64 |
| 3 | Do you complement TTI practices with other tools? | Randomized multiple choice with non-exclusive answers and other option | 62 |
| 4 | Problem(s) faced? | Open-ended | 61 |
| 5 | Recurring themes? | Open-ended | 60 |
| 6 | Prompt specifiers discovered online? | Open-ended | 57 |
| 7 | Age? | Multiple choice with exclusive answers | 64 |
| 9 | Gender? | Open-ended | 60 |
| 9 | Profession or field of activity? | Open-ended | 61 |
| 10 | A prompt you are proud of? | Open-ended | 27 |

**Table 4. Questions' types and number of respondent**

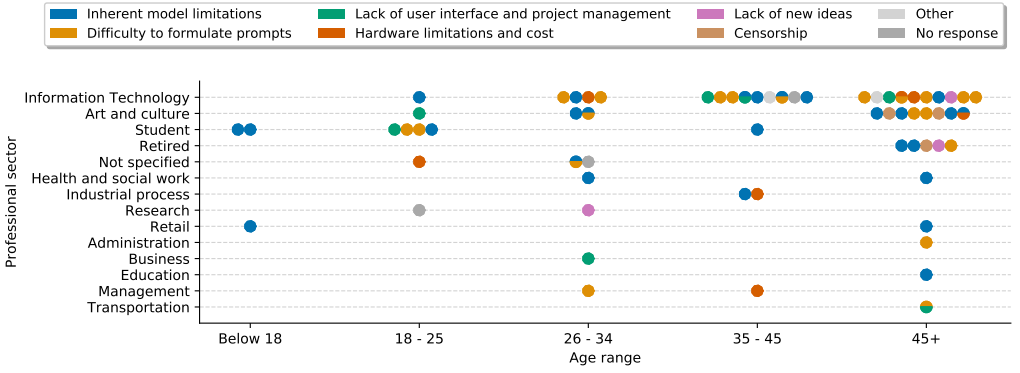
Received the 30th of january 2022

## B TEXT-TO-IMAGE USAGE AND SOCIO-DEMOGRAPHIC FACTORS


(a)




(b)




(c)

**Fig. 12. Individual representations according to age ranges (x-axis) and professional groups (y-axis) of (a) the motivations for using text-to-image generators (described in section 4.4.1, Figure 3); (b) the complementary techniques to text-to-image (described in section 4.4.2, Figure 4); and (c) the problems faced by text-to-image practitioners (described in section 4.4.4, Figure 5).**