

3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows

Vivian Liu*

vivian@cs.columbia.edu

Autodesk Research

Toronto, Ontario, Canada

George Fitzmaurice

george.fitzmaurice@autodesk.com

Autodesk Research

Toronto, Ontario, Canada

Jo Vermeulen

jo.vermeulen@autodesk.com

Autodesk Research

Toronto, Ontario, Canada

Justin Matejka

justin.matejka@autodesk.com

Autodesk Research

Toronto, Ontario, Canada

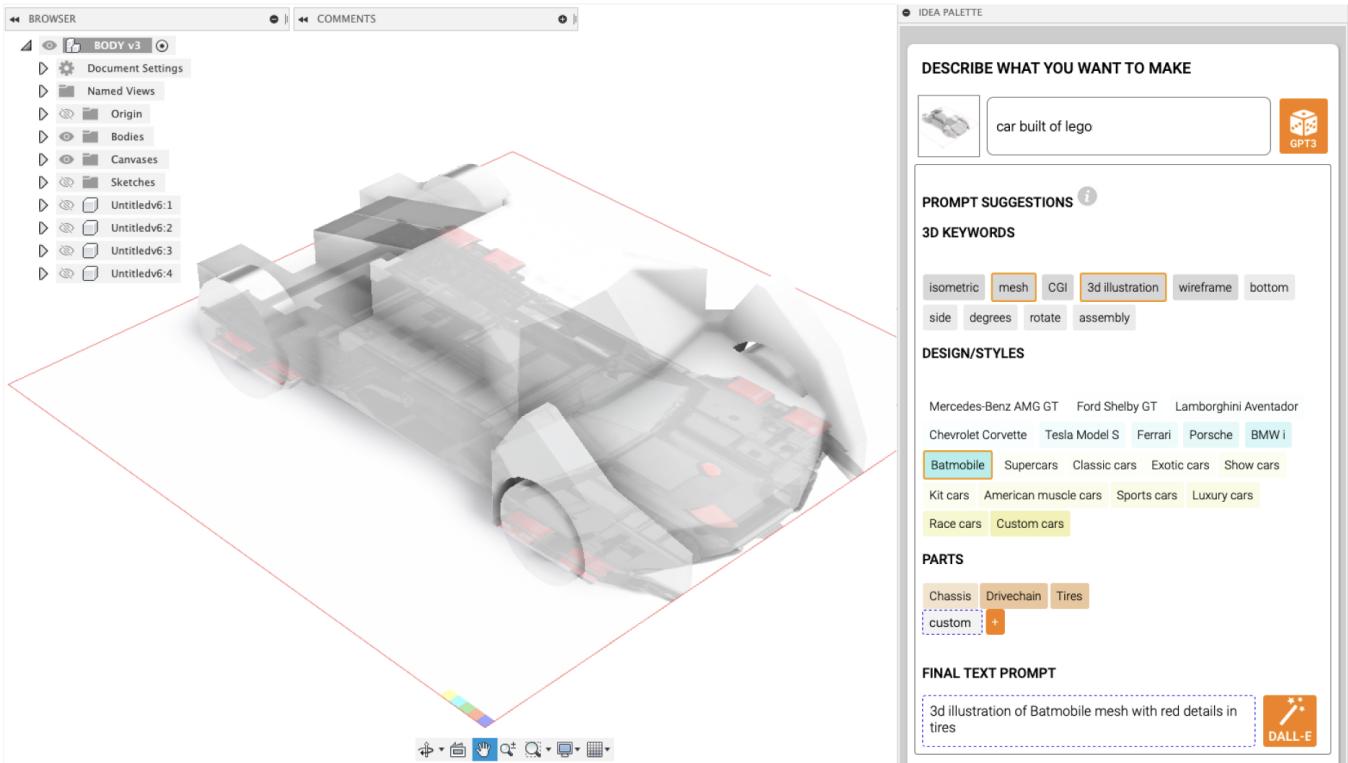


Figure 1: 3DALL-E integrates a state-of-the-art text-to-image AI (DALL-E) into 3D CAD software Fusion 360. This plugin helps users efficiently create text prompts by providing relevant 3D keywords, design/styles, and parts from GPT-3. Users can also craft image prompts by taking actions within their viewport, which can help the AI build from what a user is working on.

ABSTRACT

Text-to-image AI are capable of generating novel images for inspiration, but their applications for 3D design workflows and how designers can build 3D models using AI-provided inspiration have not yet been explored. To investigate this, we integrated DALL-E, GPT-3, and CLIP within a CAD software in 3DALL-E, a plugin that allows users to construct text and image prompts based on what they are modelling. In a study with 13 designers, we found that designers saw great potential to incorporate 3DALL-E into their workflows and to use text-to-image AI for reference images, renders, materials, and design considerations. We additionally elaborate on prompting patterns and provide measures of prompt complexity

*Also with Columbia University.

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

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

observed across participants. From our findings, we discuss how 3DALL-E can merge with existing generative design workflows and propose prompt bibliographies as a form of human-AI design history.

CCS CONCEPTS

- **Applied computing → Media arts; Computer-aided design;**
- **Human-centered computing → Interactive systems and tools;**
- **Computing methodologies → Natural language processing; Computer vision tasks.**

KEYWORDS

creativity support tools, 3D design, DALL-E, GPT-3, CLIP, 3D modelling, CAD, co-creativity, creative copilot, ideation, prompt engineering, multimodal, text-to-image, AI applications, text-to-3D, workflow

ACM Reference Format:

Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2018. 3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Designing 3D models is challenging—designers have to satisfy a number of objectives that can range from functional and aesthetic goals to feasibility constraints. Coming up with ideas takes a lot of exploration, even for expert designers, so they often consult external resources for inspiration on how to define their geometry. They browse 3D model repositories [24], video tutorials, and image search engines to understand conventional designs and different aesthetic details, trying to design in a way that is up to standard but also differentiable from the rest. However, following along with a video tutorial, using existing designs as reference images [67], and remixing 3D models [9, 24] found online only helps designers react to things that have already been made.

A recent innovation that can more directly provide inspiration to designers is text-to-image AI. Tools such as DALL-E [18], Imagen [59], Parti [69], and Stable Diffusion [58] are AI tools that have the generative capacity to access and combine many visual concepts into novel images. Given text prompts as input, these tools can capture a wide variety of subjects and styles [42]. In online communities, users have already developed a number of different methods to elicit images with 3D qualities [50]. Recent advancements have also allowed users to interact with text-to-image AI systems by passing in image prompts, where images are used as prompts in addition to text. Generations can now vary or build off of previous generations. These innovative functions make integration of text-to-image AI within existing creative authoring software more feasible.

In this paper, we seek to understand how text-to-image AI can assist 3D designers and where in creative workflows designers can most benefit from AI assistance. Furthermore, we investigate how text-to-image tools respond to image prompts sent from 3D designers as they build up complexity in their designs. To do so, we integrated three large AI models—DALL-E, GPT-3, and CLIP—with

Fusion 360, an industry standard software for computer-aided design (CAD). We implemented a plugin within the software which we call *3DALL-E*. This plugin helps translate designer's intentions into multimodal (text and image) prompts which can produce inspiration and reference images for them. After a designer inputs their goals (i.e. to design a "truck"), the plugin provides a number of related parts, styles, and designs that help users craft text prompts. These suggestions are drawn from the world knowledge of GPT-3 [5] to help users familiarize themselves with relevant design language and 3D keywords. The plugin interactively updates an image preview that shows an image prompt which can be passed into DALL-E [56], giving users a direct bridge between their 3D design workspace and an AI model that can generate related inspirational images. Additionally, having a lens on what the designer is actively working on allows the plugin to highlight what prompt suggestions may work best, which is implemented in the system by using CLIP [55] as a proxy for model knowledge. To evaluate the different parts of 3DALL-E and how well 3DALL-E can integrate into a variety of creative workflows, we conducted a user study with thirteen users of Fusion 360 who spanned a variety of backgrounds from industrial design to robotics.

We present the following contributions:

- 3DALL-E, a plugin that helps users craft text prompts with design language (different parts, styles, and designs for a 3D object) and image prompts connected to their progress
- Exploratory user study (n=13) demonstrating text-to-image AI use cases in 3D design workflows and an analysis of prompting patterns and prompt complexity

In our discussion, we propose prompt bibliographies, a concept of human-AI design history to track inspiration from text-to-image AI. We conclude on how text-to-image AI can integrate with existing design workflows and what can be best practices for generative design going forward.

2 RELATED WORK

2.1 Prompting

Prompting is a novel form of interaction that has come about as a consequence of large language models (LLMs) [5]. Prompts allow users to engage with AI using natural language. For example, a user can prompt an AI, "What are different parts of a car?" and receive a response such as the following, "Wheels, tires, and headlights". These prompts give LLMs context for what tasks they need to perform and help end users adapt the general pretraining of LLMs for these tasks without finetuning [4, 57]. By varying prompts, users can query LLMs for world knowledge, generative completions, summaries, translations, and so forth [5, 43]. Datasets around prompting are also beginning to emerge to benchmark generative AI abilities. PARTI [69] provides a schema and a set of prompts to investigate the visual language abilities of AI. Coauthor [41] provides a dataset of rich interactions between GPT-3 and writers. Audits of models have also been performed by collecting generated outputs of AI models at scale and conducting annotation studies, as in [42] and [52]. As generative AI communities have gained momentum online, crowd-sourced efforts on Twitter and Discord have also organized to disseminate prompting guidance [50].

Recent research directions have begun to create workflows around prompts. AI Chains [65] studied how complex tasks can be decomposed into smaller, prompt-addressable tasks. Promptchainer [64] unveiled an editor that helps visually program chains of prompts. Prompt-based workflows were explored in [37] to make prototyping ML more accessible for industry practitioners. Other systems have tested pipelines that concatenate LLMs with text-to-image models. In Opal [43], a pipeline of suggestions from GPT-3 initiated streams of generations from VQGAN+CLIP. Similarly, a visual concept blending system in [27] used BERT [21] to surface shape analogies and prompt text-to-image AI for visual metaphors.

New modes of prompting have also started to emerge. Users can now pass in image prompts and have AI models autocomplete images and canvases in methods called inpainting and outpainting [18, 49]. These functions have been implemented within state-of-the-art text-to-image AI systems such as DALL-E. Though some work exists on how image prompts can seed generations [52], image prompts have been less explored than text prompts. To our knowledge, 3DALL-E is one of the first to systematically create image prompts to use in a text-to-image setting.

2.2 Generative Models

Generative AI models have long been excellent at image synthesis. However, many early models were class-conditional, meaning that they were only robust at generating images from the classes they were trained on [39, 40, 53, 54, 66, 68]. The most recent wave of generative AI models can now produce images from tens of thousands of visual concepts due to extensive pretraining. CLIP [55], a state-of-the-art multimodal embedding, was trained off of hundreds of millions of text and image pairs, giving it a broad understanding of both domains. The pretraining of CLIP has also helped it serve as an integral part of multiple generative workflows [1, 15, 16, 23] and training regimes [46, 60]. Large open-source efforts had previously paired CLIP with GAN models, using it as a discriminator to optimize generated images toward text prompts. Before the popularity of text-to-image tools, practitioners of AI art were constrained to computational exploration of latent spaces [17, 30].

The novelty of generating media through language has brought many text-to-image tools into production such as Midjourney, DALL-E, and Stable Diffusion. DALL-E [56] demonstrated how CLIP embeddings can help generate images with autoregressive and diffusion-based approaches. Diffusion is key within many of the aforementioned methods to increase the quality of text-to-image outputs [14, 18, 47]. New text-to-image approaches have led to more diverse methods of user interaction. Make-a-Scene [25] allows users to interact with generations by manipulating segmentation maps, and DALL-E gives users the ability to paint outside the edges of an image, making unlimited canvases [49]. Stable Diffusion [58] gives users the ability to train and trade concepts learned by the AI. These models have extraordinary generative capacity, but their ability to be used nefariously has also inspired new approaches to safeguarding AI outputs from redteaming [6] to large scale audits for social and gender biases [12]. With 3DALL-E, we present methods to narrow the scope of general-purpose generative models (GPT3, DALL-E), such that they are useful in a specific domain: 3D design.

2.3 Creativity Support Tools

Human-computer interaction research on creativity support tools has long showcased ways to facilitate text-based content creation. Early systems showed that users could iteratively define images based on chat and dialogue [22, 61]. A number of text-to-3D methods have also been proposed; AttriBit [8] allowed users to assemble 3D models out of parts matched on affective adjectives. Sceneseer [7] and Wordseye [13] allowed users to create scenes via sentences. However, since the advancement of AI tools, much of the momentum has now concentrated around human-AI co-piloted experiences. Systems such as Opal [43], Sparks [29], FashionQ [36], and the editors in [62] are examples of AI-assisted ideation. In tandem, many frameworks for computational creativity [44] and human-AI interaction [2] have cropped up to understand concerns such as ownership and agency when AI are involved in the creative process. Gero et al. [28] found that users can establish better mental models of what AI can and cannot do if they have a sense of its internal distribution of knowledge. Cuing users towards the most helpful actions they can take within a user interface (as done in [11, 51]) is something we implement within 3DALL-E. Other best practices for creativity support tools that we revisit from an AI perspective include the idea of timelines and design history [31], natural language exploration [26], and collaboration support [63]. DataTone [26] demonstrated how interactive prompting with widgets can help build specificity in a text-based interface. Suh et al. [63] demonstrated that AI generated content could facilitate teamwork within groups by helping establish common ground between collaborators. While many systems have now been built with generative AI capabilities [19, 30, 45] and even for text-to-image workflows [43]—none that we know of have applied text-to-image AI for 3D design workflows.

3 DESIGNING WITH 3DALL-E

3DALL-E is provided as a panel on the right hand side of the 3D workspace (Fig. 1). Fig. 2 shows the steps users go through when designing with 3DALL-E inside their 3D workspace and presents the main interface components. 3DALL-E allows users to construct prompts relevant to their current 3D design, which can then be sent to DALL-E to retrieve AI-provided image inspiration. Once generations are received, users are able to download them, see a history of previous results, and create variations of generations that they want to explore more from. In what follows, we will present these different steps with a short walkthrough.

3.1 Constructing Text Prompts for AI-Provided Inspiration

Users begin at the starting state shown in Fig. 2-I, where they can describe what they want to make by typing in their goal (Fig. 2A). Once they do that, different prompt suggestions populate the sections with 3D keywords, designs/styles, and parts (Fig. 2-II). These suggestions help steer the generations toward results relevant to 3D modeling as well as provide design language a user might otherwise not be familiar with. For example, querying a chair could return a series of existing designs such as an egg chair, an Eames chair, or a Muskoka chair, helping familiarize the user with the design

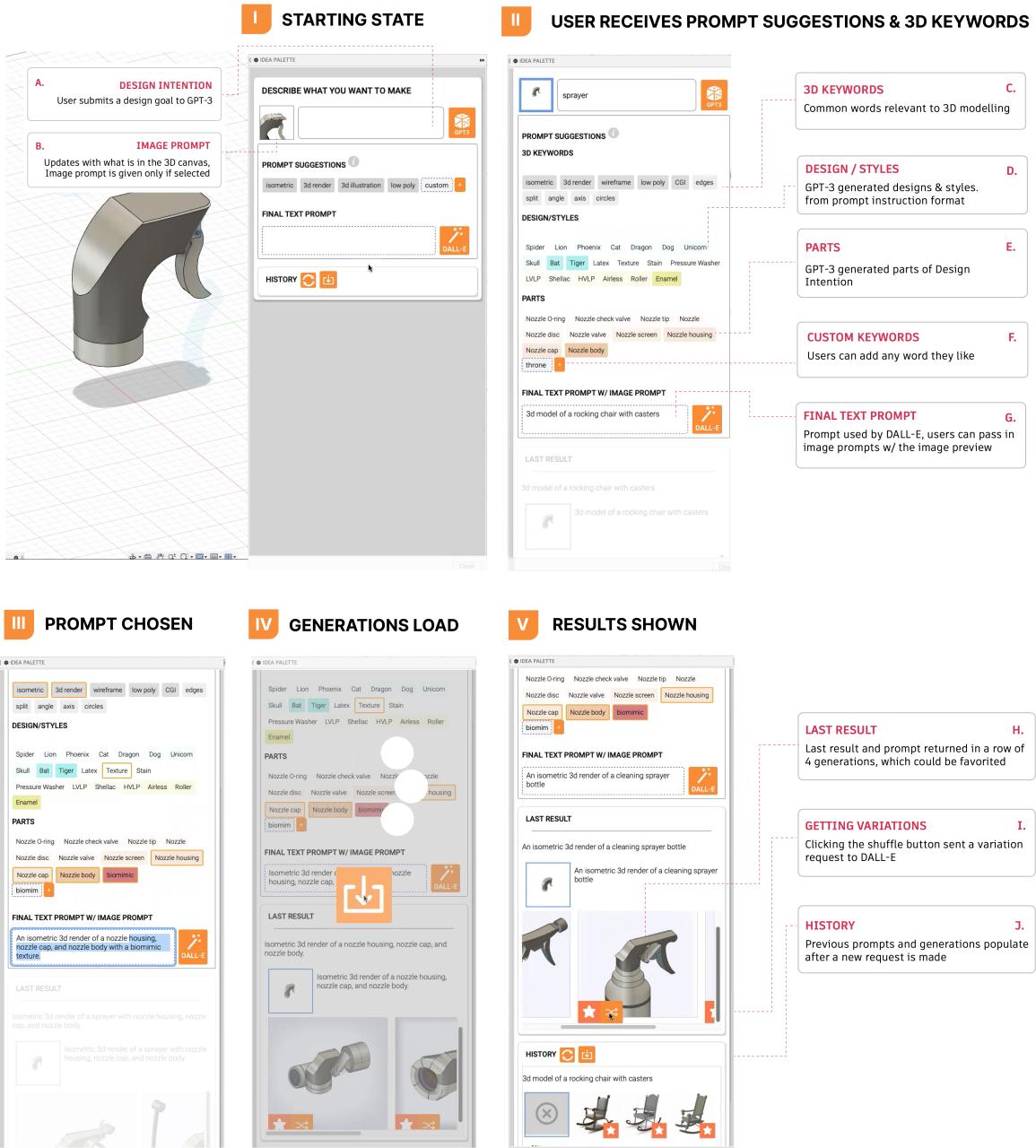


Figure 2: 3DALL-E walkthrough. Step I: Initial state, where users can type their design intentions. Step II: Users are presented with prompt suggestions from GPT-3. Step III: Selected suggestions are rephrased into an editable prompt. Step IV: Users wait as DALL-E generates. Step V: Results are shown. A cursor hovers over a shuffle icon, which is how users can launch variation requests from DALL-E.

language befitting of chairs. Once users select a set of prompt suggestions (e.g. “*3d render, isometric, plant stool, wrought-iron*”), an automatically rephrased prompt appears in the final prompt box (e.g. “*isometric 3d render of a wrought-iron plant stool*”) as shown in Fig. 2-III. This prompt is still editable by the user, and a text box to

add custom keywords is also available when clicking the orange ‘+’ button in the parts section (Fig. 2F).

Prompt suggestions (Fig. 2C–E) are color-coded with a color for the group they belong to (blue for designs, green for styles, orange for parts) and varied in opacity to indicate how strongly their text

aligns with the image prompt (see Fig. 4 for implementation details). For example, from a set of styles like “mid-century modern, contemporary, and art deco”, if “art deco” was most strongly highlighted (i.e. more opaque – darker green), it meant that the image prompt had the greatest probability of being matched with “art deco”.

3.2 Crafting an Image Prompt

Users can also choose to include an image that is automatically extracted from their current 3D modeling workspace in addition to their text prompt (image+text prompt) or choose to exclude it (text-only prompt). Image prompts are only passed in when users select the image preview (Fig. 2B), making it active. Using the 3D software to render the viewport allows 3DALL-E to programmatically deliver clean prompts without tasking the user with any erasing or masking. Users can easily toggle the visibility of certain parts of their model using the 3D software’s built-in functionality and request for DALL-E to fill in the details for those hidden parts.

3.3 Receiving Results from DALL-E and Retrieving Variations

Once the user is satisfied with the prompt, they click the DALL-E button next to the final text prompt (Fig. 2G) to generate either a text-only or image+text prompt (depending on whether the image preview is selected). While waiting for results (Fig. 2-IV), the user is shown a spinner animation. When the results are ready, the user can click the orange download button to pull the results from DALL-E into the 3DALL-E interface.

Results are returned in sets of four (Fig. 2-V). When the user hovers over a result, they are presented with a menu that allows them to ‘star’ their favorite results and click the ‘shuffle’ button to get more *variations* on that particular result (Fig. 2I). Lastly, 3DALL-E also keeps a history of previous generations (Fig. 2J).

4 SYSTEM IMPLEMENTATION

3DALL-E was implemented within Autodesk Fusion 360 [33] as a plugin and written with the Fusion 360 API, Python, Javascript, Selenium, and Flask. Fig. 3 illustrates how we embedded DALL-E, GPT-3, and CLIP into one user interface. All actions in 3DALL-E were logged by the server to facilitate analysis of participant behavior in the study (Sect. 5). Note that 3DALL-E could be implemented generically in most 3D modelling tools. The needed functionality from Fusion 360 is relatively basic: a custom plugin system, ways to render the viewport as images and to apply textures to models.

Prompt suggestions were populated by querying the GPT-3 API for the following: “List 10 popular 3D designs for {QUERY}? 1.”, “What are 10 popular styles of a {QUERY}? 1.”, and “What are 10 different parts of a {QUERY}? 1.” To rephrase chosen suggestions, GPT-3 was prompted: “Put the following together: {SUGGESTIONS}”.

Ten 3D keywords are sampled from a set of high frequency words ($n=121$) in a Fusion 360 Screencast dataset. Screencasts are videos used to communicate help and tutorials in forums [3, 31]. Automatic speech recognition (ASR) of these videos produced transcripts; these transcripts were processed with natural language processing, filtered out for general purpose words, and sorted by frequency to get the final keywords set.

Text highlights were calculated by passing each of the prompt suggestions and the image prompt to CLIP, which was hosted on a remote server. CLIP produces logit scores that suggest how similar each text option was to the image, a value 3DALL-E renders as the opacity of each highlight. The stronger the highlight, the greater the probability a text option matched what a user had in their viewport. DALL-E was trained with CLIP text and image embeddings. By using CLIP’s embedding in this way, users receive a computational guess for how well DALL-E might be able to interpret each prompt suggestion, while also dialing down the options they need to focus on (Fig. 4). The 3D keywords were by default gray, while designs, styles, and parts were matched to gradations of blue, green, and yellow respectively.

We used the Fusion 360 API to automatically save the viewport to a PNG image every 0.3 seconds. The workspace of Fusion 360 (the gridded background pictured in Fig. 1) was rendered transparently in the PNG image.

5 EVALUATION

Implementing 3DALL-E within Fusion 360 gave us a focused application context to evaluate text-to-image AI within a creative workflow. We set out to investigate the following research questions for 3DALL-E—all of which fall under a main question: **are text-to-image AI systems useful within 3D design workflows?**

- *Prompt patterns.* Are there certain patterns of prompting that can be observed?
- *Highlighting model knowledge.* Does adding approximate cues for what text-to-image AI might know improve user experience?
- *Prompt complexity.* How many concepts do people like to put within prompts?

To do so, we conducted an exploratory study with 3D CAD designers ($n=13$, 10 male, 3 female). Participants were recruited from internal channels within a 3D design software company as well as through a design institute mailing list at a local university. Participants were compensated with \$50 dollars for 1.5 hours of their time. The average age of the participants was 28, and they had an average of 4.13 years of experience with Fusion 360 (min=1 year, max=8 years). Five had experience with the generative design environment within Fusion, and three had prior experience with AI / generative art systems. The participants spanned a range of disciplines from machining to automotive design. Domains of expertise and frequency of use of the 3D software are listed in Table 1.

5.1 Experimental Design

Participants were given two tasks: T_{edit} to edit an existing model and T_{create} to create a model from scratch. The intention of having these two tasks was to show how 3DALL-E might affect creative workflows at different stages of the 3D modelling process. The ordering of these tasks was counterbalanced to mitigate learning effects. This experimental design was approved by a relevant ethics board.

Before the study, participants were sent an email with DALL-E’s content policy to disclose that they were going to use AI generative tools. During the study, participants were given a brief introduction

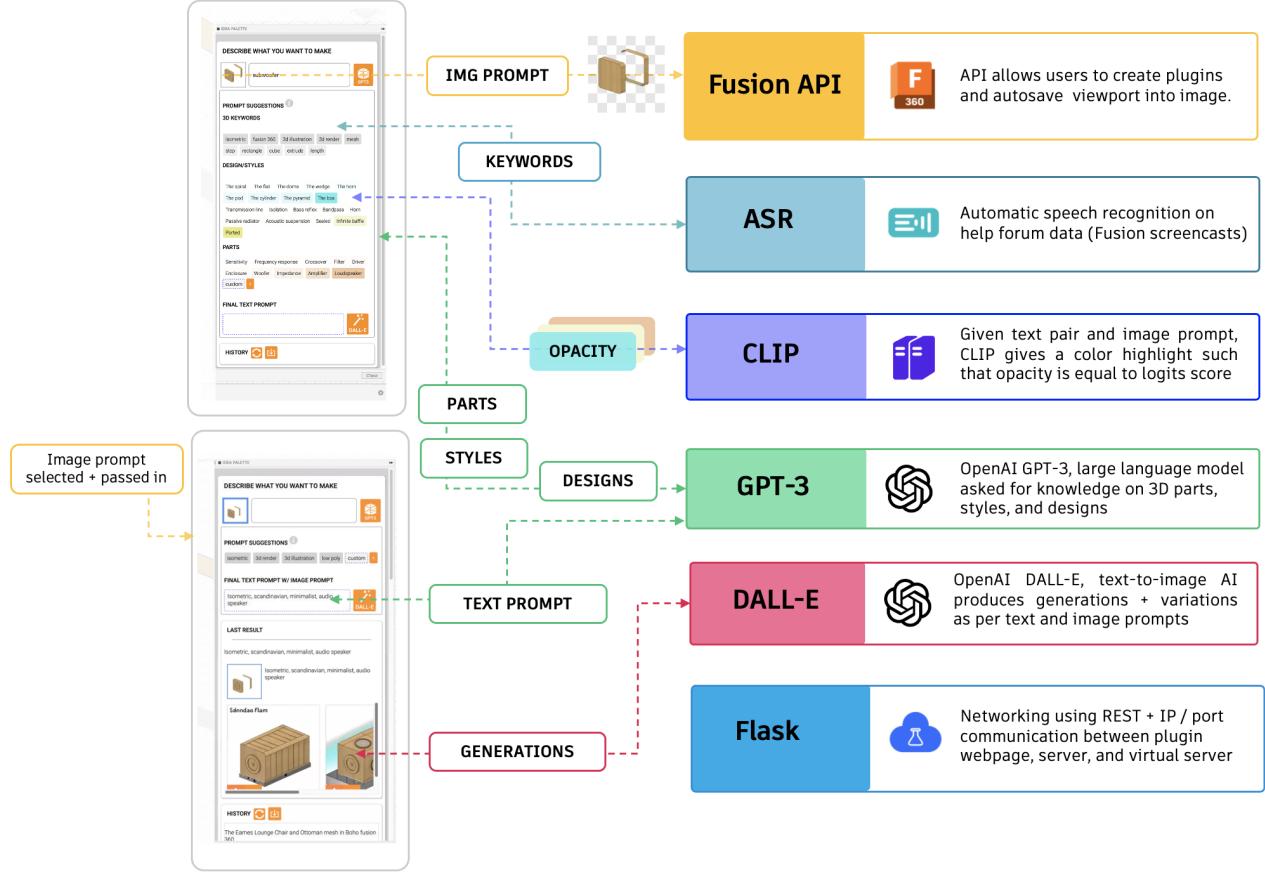


Figure 3: System design showing the architectures involved in 3DALL-E, which incorporates three large AI models into the workbench of an industry standard CAD software. In the top left panel, we show how text AI outputs are displayed in the UI. In the bottom left panel, we show how users could pass in image prompts and retrieve DALL-E generations within the plugin.

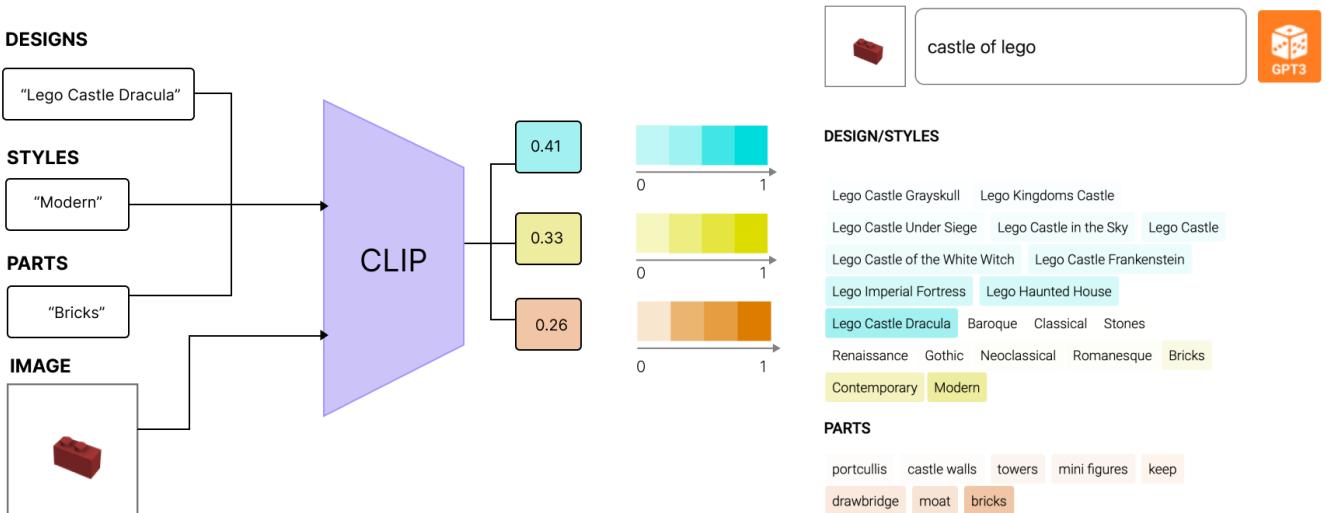


Figure 4: Diagram showing how text highlights were calculated using CLIP with image and text from the prompt suggestions as input. The CLIP logits score was set as the opacity of each prompt suggestion. Each type of suggestion was colored differently.

to the different AI architectures involved (GPT-3, DALL-E) and given two general tips on prompting: 1) text prompts should include visual language, 2) text prompts are not highly sensitive to word ordering [42]. Participants were then given a walkthrough of the user interface and the different ways they could generate results from GPT-3 and DALL-E. The study was conducted virtually via Zoom and through remote control of the experimenter's Fusion 360 application and plugin.

T_{edit} was to modify an existing 3D model that the participants had brought with them to the study. Participants were told to bring a non-sensitive model, meaning one that did not include corporate data. There were no constraints on what the model could have been. Examples of models brought in can be seen in Fig. 5. When a participant did not have a model to use, a random design was provided from the software's example library. This was the case for only one participant (P15).

For T_{create} , participants were allowed to pick whatever they wanted to design from scratch. For each task, participants had thirty minutes to work on their model with the assistance of 3DALL-E. At the halfway point, participants were reminded of the time remaining and of any generation actions that they had not tried out yet from GPT-3 (prompt suggestions) or DALL-E (text-only prompts, image+text prompts, variations). Beyond this reminder, they were guided only if they needed assistance accomplishing something in the user interface. Examples of what participants created for T_{create} can be seen in Fig. 6. At the thirty minute mark, designers were told to wrap up their last design actions.

After completing each task, participants marked generations in their history that they felt were inspiring and completed a post-task questionnaire, which included NASA-TLX [32], Creativity Support Index (CSI) [10, 45], and workflow-specific questions. These questions can be found in the supplementary material. A semi-structured interview was then conducted to understand their experience with 3DALL-E.

5.2 Feedback on 3DALL-E

The metrics we measured showed that designers responded to 3DALL-E with enthusiasm. In terms of enjoyment, 12/13 participants rated their experience positively (≥ 5 out of 7) for T_{edit} and 11/13 for T_{create} . The majority of participants also responded positively that they were able to find at least one design to satisfy their goal: 10/13 respondents in T_{edit} , 12/13 respondents in T_{create} . We note that for T_{create} 7 of the participants responded with the maximum agreement (strongly agree). Likewise, most participants reported that the system helped them fully explore the space of designs (9/13 responded positively for T_{edit} , 11/13 for T_{create}).

"I could spend ages in this." -P18

In general, the post-task questionnaire results were similar for T_{edit} and T_{create} . However, on a few dimensions, participant responses were distributed slightly differently. For example for effort, while responses for T_{edit} about tool performance ("How successful were you in accomplishing what you set out to do?") were split across the spectrum, with 6/13 rating the tool positively, 10/13 participants rated the performance for T_{create} positively. In terms of ease of prompting, while 13/13 respondents were positive that for T_{create} it was easy to come up with prompts, 10/13 responded positively for

T_{edit} . We hypothesize that this could have been because for T_{edit} participants had to work under more constraints, bringing in 3D models that were often custom and near finished.

We note that frustration was low for both Tasks; 11/13 responded on the low side of the spectrum for T_{edit} (≤ 3), 10/13 for T_{create} . For T_{edit} , frustration was low in spite of the fact that 6/13 of participants disagreed to some degree (≤ 3) about having control over the generations they were creating.

"The amount of control you have with the system is very dependent upon how specific you get with the text. For example, if I make it super broad, you're obviously going to have less control because DALL-E is working off of less information. So it may provide its own information. It has to kind of fill in the gaps of what you're trying to say. But the more specific I got, the better results I got." -P1

"It was a bit difficult to control. Some things I wasn't quite expecting. For example, with this one [generation of a watch] I expect that it would have more circular watch faces, but it came with ones that were more angular." -P8

Lastly, we discuss workflow-specific questions about the prompting pipeline of 3DALL-E. Participants were asked about the usefulness of 3DALL-E to their usual workflow. For T_{edit} , 10/13 felt that it would be helpful. For T_{create} , 10/13 also felt it would be useful (with 7 giving 3DALL-E rating 7 for strongest agreement). In another question, we asked whether it was easy for participants to come up with new ways to prompt the system. Participants responded unilaterally positively for T_{create} (13/13 responded ≥ 5) and positively for T_{edit} (10/13 responded ≥ 5). Participants were also asked to rate how useful they found the GPT-3 suggestions. For T_{edit} and T_{create} , the responses were generally positive, at least 8/13 participants responded with 5 or higher for both tasks.

"I'm looking for the right word and I think that's where this text [GPT-3] search can come in handy... I think it's helpful to know its language, to know what it finds." -P4

"I think having the GPT-generated ones was useful. It allowed for some ideas I didn't consider... [ideas I] wouldn't have found the words for." -P13

On whether or not the highlighting of prompt suggestions was useful, participants responded with more even distributions, though the distributions still skewed positive (8/13 in T_{edit} and 7/13 in T_{create} gave the statement a 5 or higher). Lastly, we gauged participant response to image prompts, asking if they agreed that image prompts were incorporated well in their generations. For T_{edit} , 10/13 participants responded with a 5 or 6 for agreement. For T_{create} , 8/13 participants responded with a 6 or 7.

"Image prompts definitely allowed me to tailor the outcomes towards what I was hoping for or expecting maybe... I'd have struggled to replicate [the render type] if I hadn't done the click on the image [sent in an image prompt] and create some variations. I think once I found something I

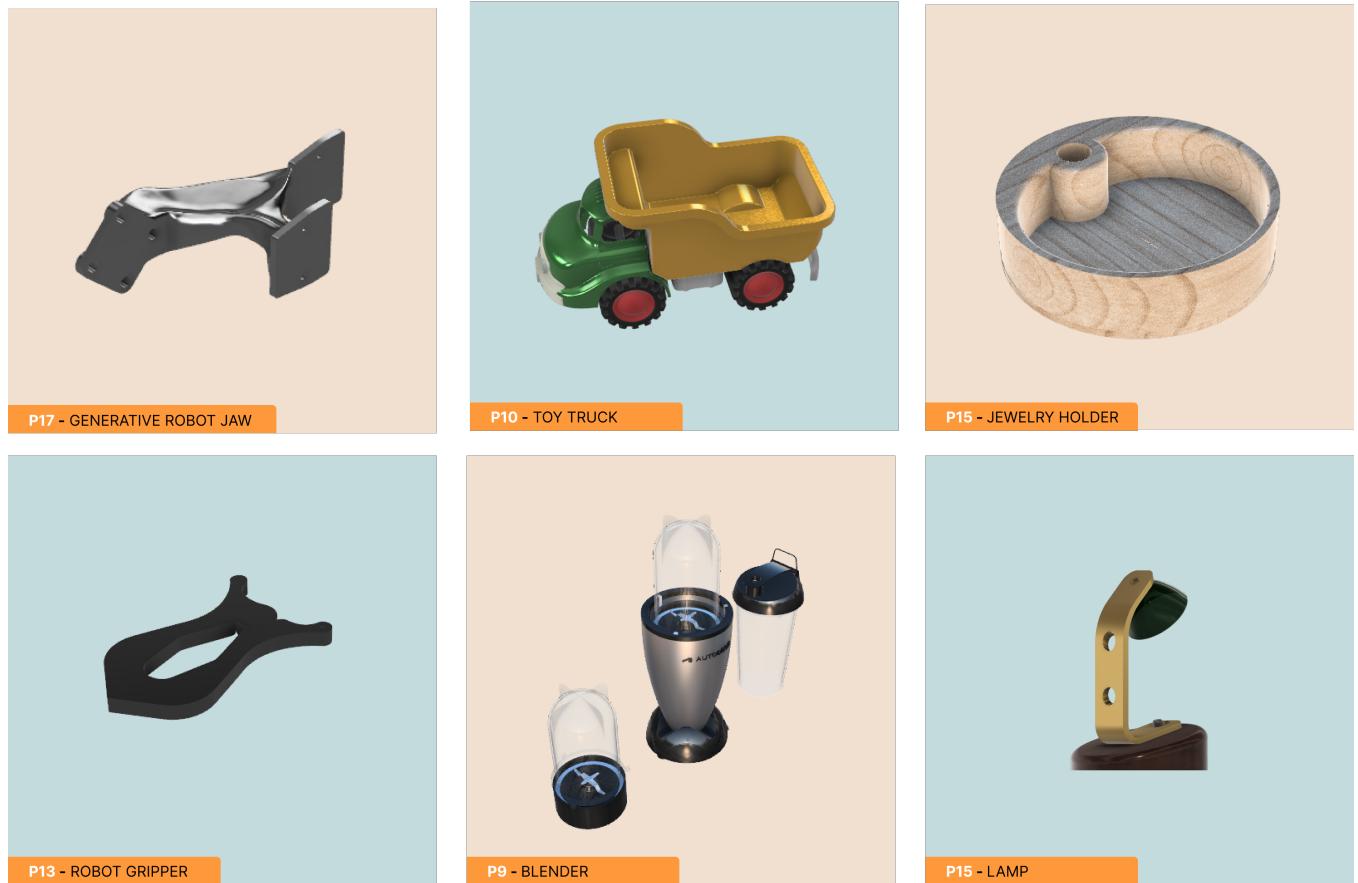


Figure 5: Examples of 3D designs participants brought in during T_{edit} , which was to edit an existing model.

Table 1: Table of participant details, with discipline and Fusion 360 usage frequency. We list labels for the model they designed during T_{create} and labels for the model they brought in (T_{edit}).

ID	Discipline	3D Modelling Freq	T_{edit}	T_{create}
P1	Mechanical engineering, robotics	Few times a week	robot	prosthetic hand
P2	Design/ drones	Daily	drone	airplane
P3	Design	Few times /year	ring	iPhone
P4	Technical sales specialist	Daily	machined part	bluetooth ear gauge
P5	Mechanical engineering	Few times/ month	jewelry holder base	outdoor 3D scene
P8	Mechanical engineering	Few times/ month	top	table
P9	Mechanical engineering	Few times /year	spray bottle mechanism	mittens
P10	Technical accounts executive	Few times /week	truck	shelf
P11	Technical support (machining)	Daily	blender	bottle
P13	Software engineer, industrial design	Daily	gripper	speakers
P15	Technical product manager	Few times a year	lamp	bookshelf
P16	Mechanical engineer	Few times a year	sensor mount	screwdriver
P18	Technical sales, industrial design	Daily	microphone stand	car

liked, using those variations made it much easier to stuck to that design theme.” - P13

“This middle one is pretty insane... it has integrated my design into the image properly... even

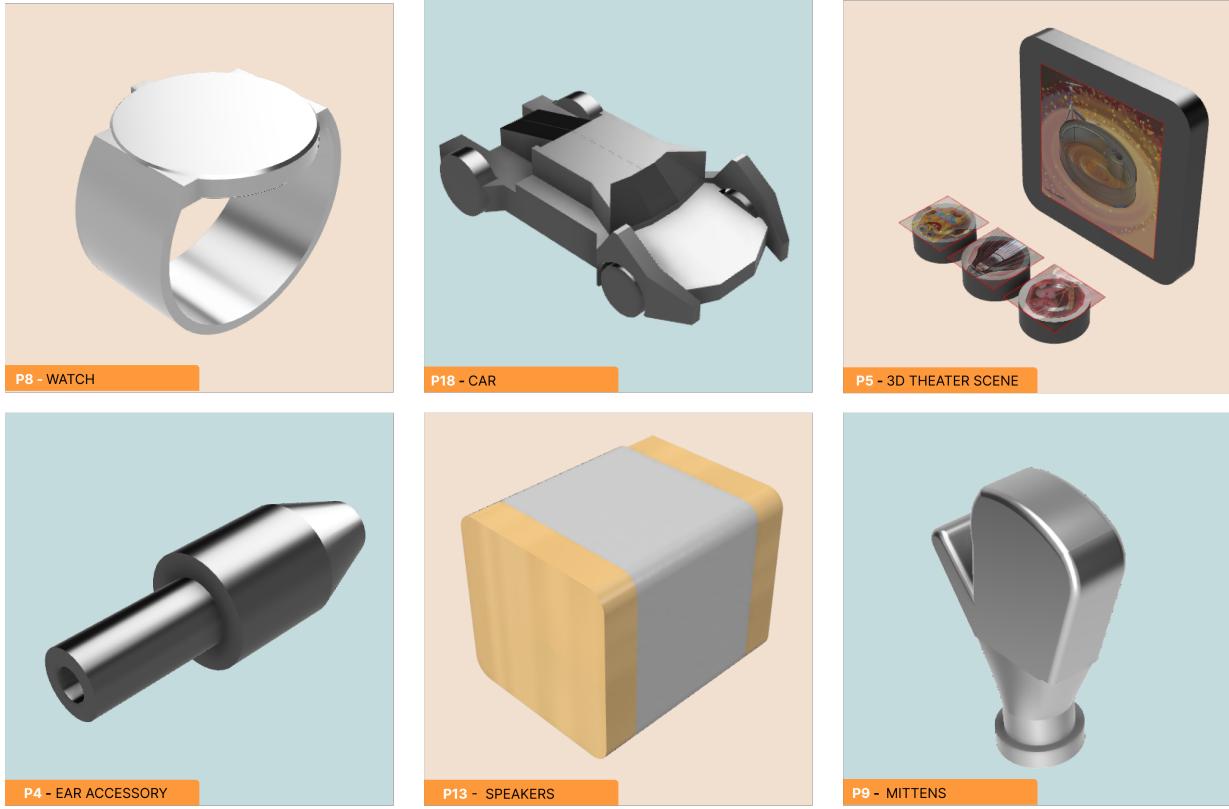


Figure 6: Example of 3D designs participants came up with during T_{create} , which was to create a model from scratch.

as an assembly, I think that's completely nuts... [An image prompt] connects what I'm working on with it [DALL-E]... otherwise it might be giving some random results, and after a while it might become redundant for me.] - P18

5.3 Prompting Behavior

We were able to observe certain patterns of prompting with 3DALL-E as each generation action was logged by our interface. From these logs for both GPT-3 and DALL-E, we were able to provide timelines of generation activity in Fig. 8 (T_{edit}) and Fig. 9 (T_{create}).

5.3.1 AI-first, AI-throughout, or AI-last. One of the most salient ways to distinguish participants was at which points in their workflow they took to 3DALL-E and which points they focused on Fusion 360. Some participants were *AI-first*, meaning they tended to sift through AI generations first until they had a better grasp of its abilities or until they found a design that they liked before taking any significant 3D design actions. For example, P18 (top row of Fig. 8), a technical software specialist with an industrial design background, first began looking for inspiration for a matchbox car, before diving into prompt suggestions like “sports car”. Text prompts that P18 tried included “*a single sports car built like a Lego building block, view from the top.*” and “*The Dark Knight Rises: the body of a car as*

a Lego building set”. After liking one of the resulting generations (Fig. 11), P18 used the result as a reference image. For the rest of the duration of the task, P18 modelled within Fusion 360. The modelling experience during T_{create} is shown in Fig. 13.

The *AI-last* pattern occurred when participants jumped straight into their existing workflows for 3D design and tried 3DALL-E later. We see this in the rows of Fig. 9 that start off with orange bars, which indicate that participants started modelling from the get-go of the task. P11, for example, began by creating a bottle out of simple sketches and extrusions. Only after they had created a basic model did they start looking for inspiration; seeing generations of Coca Cola bottles *later* helped P11 figure out how to bring complexity into their design. P16 (second row in Fig. 9) was another AI-last participant. They already had an existing screwdriver concept in their mind. Only once that concept was in the canvas did they begin to query the model to see if they could get DALL-E to replicate their 3D design. Note that the *AI-last* pattern, jumping into a participant’s existing workflow with Fusion 360, was more prevalent in T_{create} .

However, there were also participants who queried *AI-throughout*. Many participants (P13, P1, P8, P10) would intermittently craft an image prompt by briefly working within Fusion 360 and then start generating. We see these actions whenever participants would have a short window of Fusion time that led up to image+text generation (medium blue dots in Fig. 8 and Fig. 9). During these short windows,

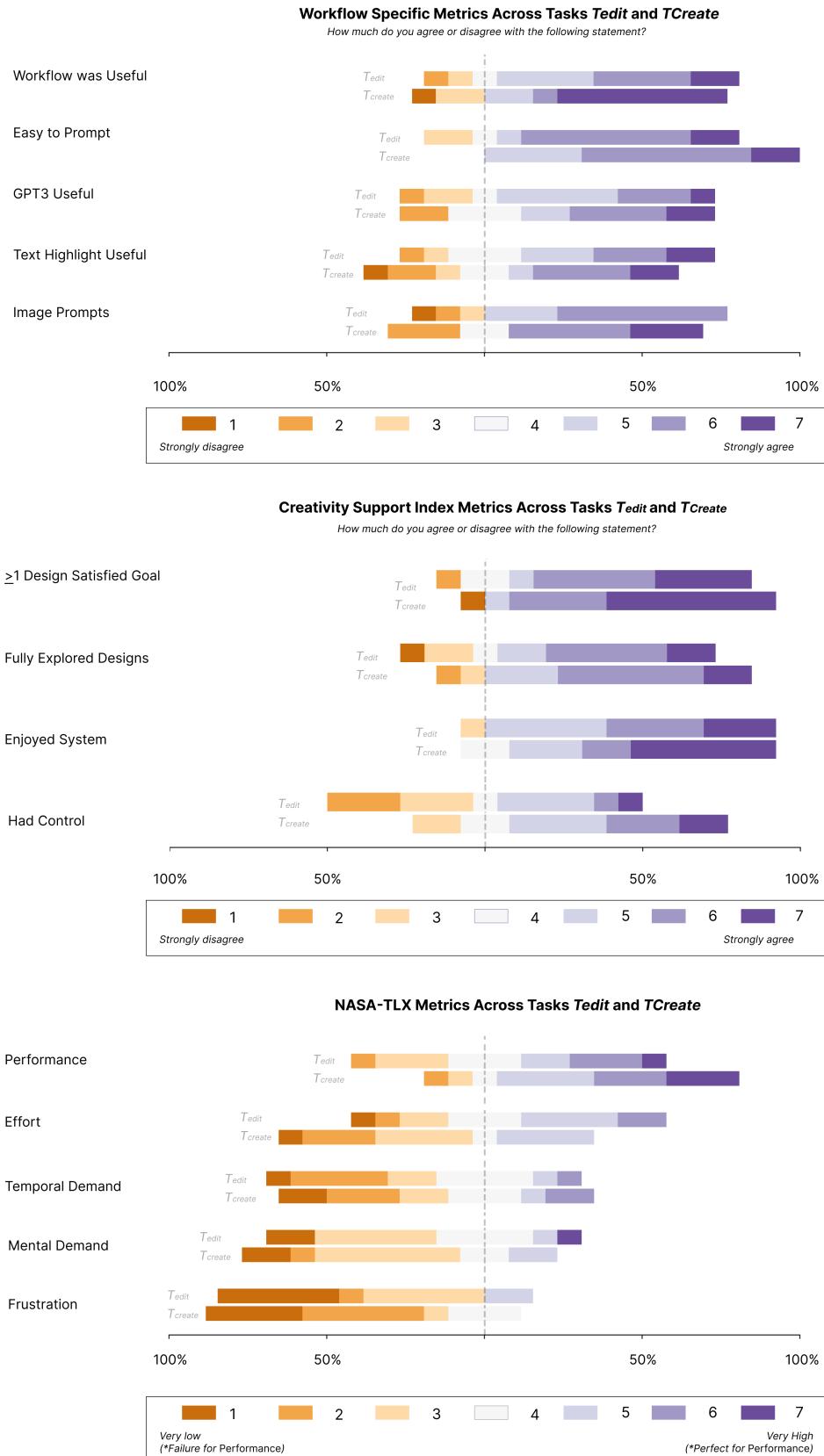


Figure 7: Distribution of Likert scale responses on NASA-TLX, creativity support index, and workflow-specific questions across all participants for both T_{edit} and T_{create} . Full questions are in the Appendix.

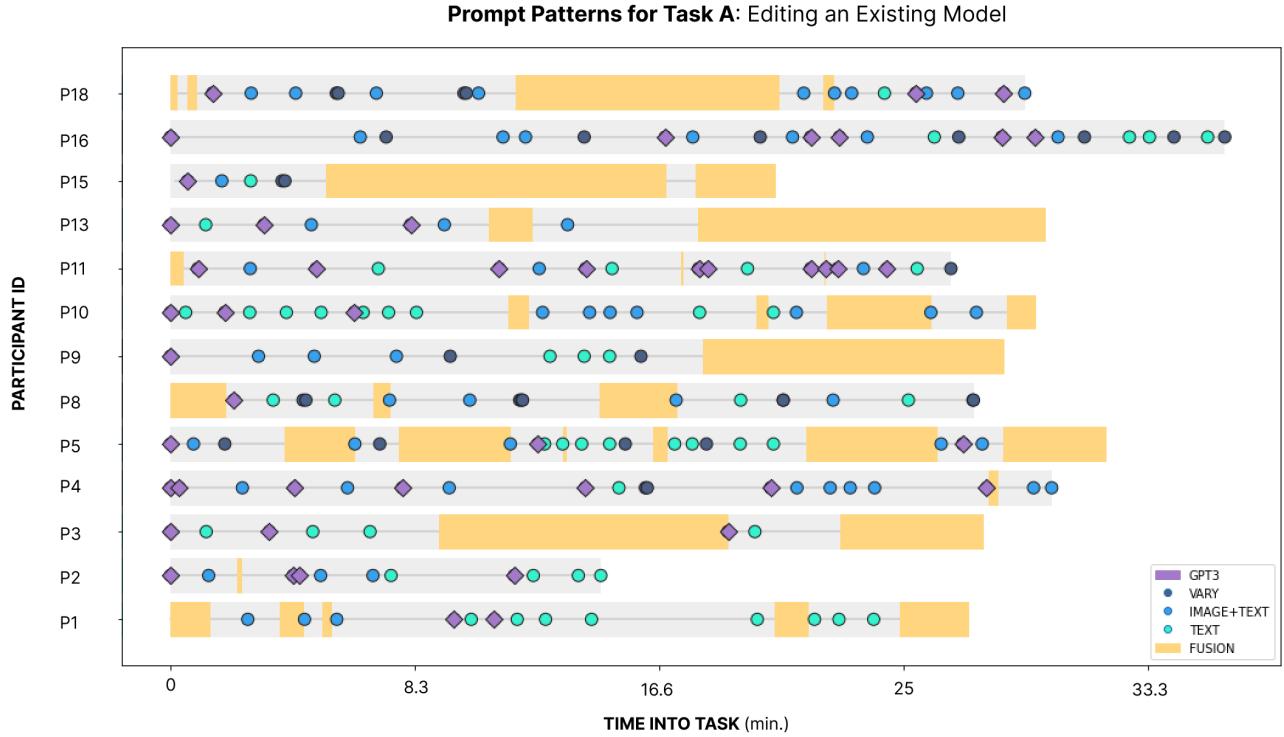


Figure 8: Pattern of generation activity for T_{edit} , when participants edited an existing model.

participants were generally changing their camera perspective or the visibility of different parts in their assemblies. For example, P10 hid the hopper of a toy truck they had brought in and tried to generate different semi-trailers using prompts such as “*Jeep Gladiator snow plow truck*”. P13 went a step further and performed destructive operations on their model geometry, deleting faces and extrusions in their geometry to get the right image prompt they wanted.

Participants would also use text-only prompts to take them towards new directions. P9 used text prompts to pivot their design multiple times and better scope their 3D design. Originally, P9 intended on creating a prosthetic hand and tried generating “*A 3D model of a robotic hand with two fingers*”. After finding this too challenging, they pivoted twice more (from a 3D fist to a mitten), each time generating from text-only prompts to explore a new direction.

In terms of generation patterns for GPT-3, nearly everyone started with generating from GPT-3, (though this could be because of the organization of the user interface). Many continued to use GPT-3 throughout each task, and we can see this reflected in the fact that there are purple diamonds, indicating GPT-3 action at the early, middle, and late stages of participants for both T_{edit} and T_{create} .

5.3.2 Switches between Types of Prompting. Eight participants passed in *an image prompt* as their first action in T_{edit} , and eight participants passed in *text prompts* as their first generation action for T_{create} . This suggests that participants may be more likely to pass in an image prompt if they already have work on their page.

Aggregating across all the different generations across T_{edit} and T_{create} , we did not see that any mode of prompting was favored more than the rest. Preferences in prompting were highly dependent upon the participant and also how well the participant felt like the generations incorporated their image prompts. For example, even though P13 found image prompts useful, P13 felt like image prompts were incorporated in an “awkward” way, as they had more glaring visual artifacts than generations from text-only generations or variations. Others like P4 were delighted by the way the visual artifacts could take their models in unexpected directions. For example, they greatly enjoyed how DALL-E produced a kaleidoscopic image out of their machined part.

In certain rows in Fig. 8 and Fig. 9, we could see that some participants would shift away from using image prompts and focus on text-only prompts. A case in point of this was when P1 worked on a tank-drive robot during T_{edit} . To craft image prompts, they played around with different angles of their models and toggled the visibility of parts like the wheels and ground plane of their model. While they found that 3DALL-E could generate decently even on visually complex image prompts, they ended up passing in a series of text-only prompts like “*3D illustration of a Roomba with four wheels powered by motors*” and “*flat image of a toy wheel*”. (Focusing in on a specific part rather than trying to get 3DALL-E to work with the full assembly was also a common strategy of participants.) In this situation, the text-only generations were easier for P1 to parse and make sense of. P5 was another example of someone who pivoted

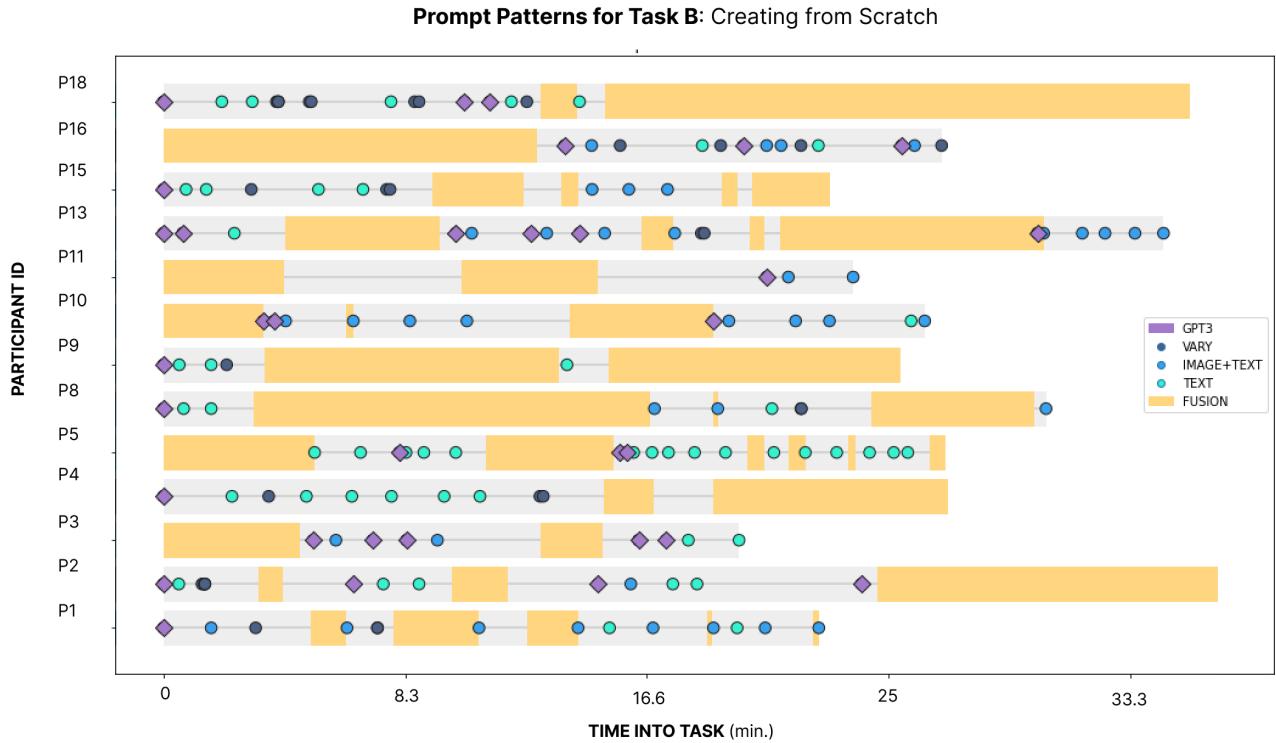


Figure 9: Pattern of generation activity for T_{create} , when participants created a model from scratch.

away from passing in image prompts to use text-only prompts after receiving sets of unsatisfying generations during T_{edit} . They instead decided to generate textures of water and maple syrup to project onto their original model, finding the resulting generations less abstract. People also tended to utilize text-only prompts when they wanted to pivot design directions. P8 used text-only prompts to try different design styles (from “industrial minimalist” to “traditional farmhouse” to “nature-inspired Scandinavian”) in quick succession.

6 PROMPT COMPLEXITY

In 3DALL-E, GPT-3 would automatically rephrase selected prompt suggestions while adding a small amount of connecting words. Based on this design, we could measure complexity as the number of concepts forming the basis of a prompt. For example, if “*3d render, minimalist, chair*” was rephrased as “*3d render of a minimalist chair*”, we gave the prompt a count of 3 concepts.

However, participants also had the ability to edit the final prompt and to add or subtract concepts of their own. In cases where the text prompt mostly came from the participant rather than GPT-3, we counted the number of concepts based on rules from linguistics and natural language processing. The prompt complexity was then the number of noun phrases and verbs in a prompt, ignoring prepositions, function words, and stop words. Count words were ignored; they were considered modifiers for the noun phrases they were a part of (e.g. “five fingers” was one concept).

We annotated text-only and image+text prompts with the number of concepts. We did not annotate variations for complexity because the generation of those images were not directly informed by text prompts. From these annotations, we charted prompt complexity across participants in Fig. 10. We found that participants tended to explore between two to six prompts, which is where most of the density of points concentrates in Fig. 10. We see that participants were also willing to try a range of concepts, as we can see in the wide spread of P2, P9, and P10. Fig. 10 also shows that participants could easily assemble prompts of over six concepts with this workflow.

We note that even when the prompts were filled with concepts: “*V-shape, Y, Tricopter, Sports, Abstract, Landscape, Aerial, Gimbal, Camera, Transmitter, Flight controller, Receiver*”, 3DALL-E could still return legible images. For this prompt, P2 received generations that had laid out displays of product components. P2 was an obvious outlier in the complexity of the prompts that they provided. They were keen on trying to “break the system” and passed prompts averaging 10 concepts. We did not discern a difference between complexity observed for T_{edit} and T_{create} .

7 QUALITATIVE FEEDBACK

7.1 Industry Use Cases

Participants demonstrated different use cases of 3DALL-E as they progressed through the tasks. The most commonly acknowledged use case was using the system for inspiration, particularly in the

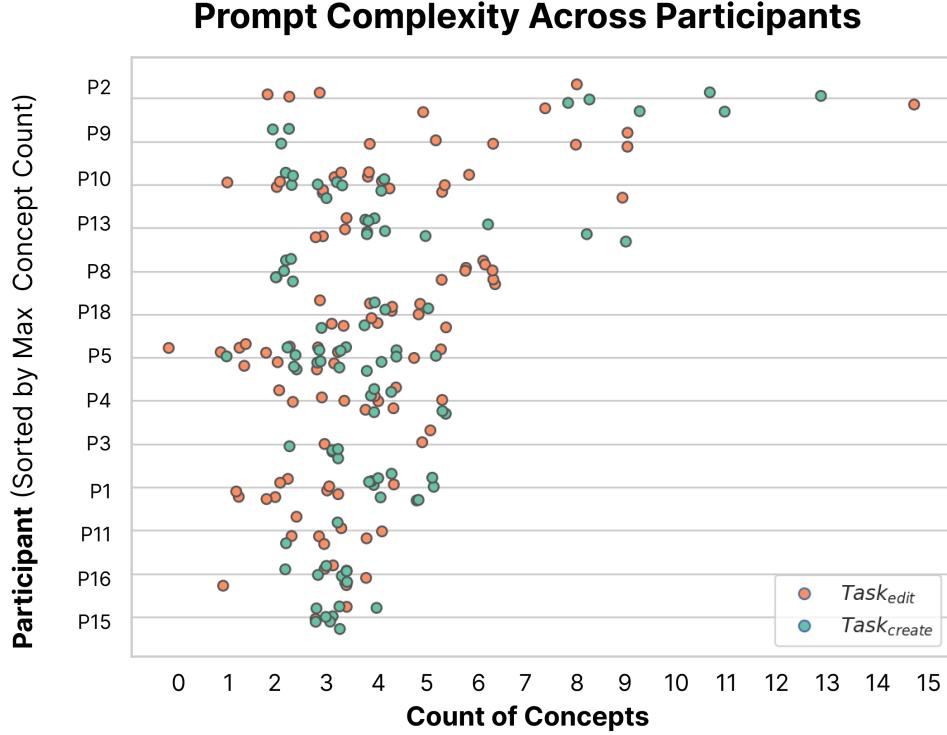


Figure 10: Prompt complexity measured across participants, where complexity is the count of concepts in each text-only and image+text prompt. Participants span the X-axis, sorted by the count of their most complex prompt. The values are jittered to show multiplicity; many prompts mapped to the same number of concepts. Complexity tended to concentrate between two to six concepts, as seen by the density of prompts within that interval. Each datapoint was colored based on the task the prompt was from.

early stages of a design workflow. P10 contextualized some of the challenges that 3D designers face on the job, such as design fixation and time constraints. “A lot of times designers get stuck, they get tunnel vision...the folks at [toy design company] used to say to me, ‘We can’t come up with enough designs’...it takes too long to come up with a design, so then we only get two or three...we would like to see thousands of design options and variations...the [designer’s] goal is to start throwing as many designs out there as they can.”

Participants felt like the tool could be “game-changing” (P11) for certain industries, particularly for designers of consumer products, automobiles, and game assets (P10, P11, P15, P17, and P18). They likened it to existing search and intelligent suggestion tools like the stock photography websites (P15) and Google Images, but noted that with 3DALL-E, it was better in that users could access inspiration without leaving their workbench (P11). The lack of application switching is important in 3D design software, which require focused modelling time. In the following sections, we elaborate use cases for text-to-image AI in 3D workflows.

7.1.1 Use Case: Reference Images for 3D Geometry. Many participants (P18, P3, P13) imported generations into the 3D software as reference images to model off of. P18 and P13, both of whom had backgrounds in industrial design, described how designers



Figure 11: Three DALL-E generations participants (P18, P9) found inspirational from the prompts: “*The Dark Knight Rises: the body of a car as a Lego building set top view*”, “*3D render of a desk lamp Victorian*”, and “*isometric 3d renders of a cleaning sprayer bottle*”.

traditionally gather reference images to build their models. These images generally aligned with specific views: front, side, top-down, perspective, or isometric. P18 said, “I would probably need at least three images: top, side, and front view to even understand it three-dimensionally...that’s what a designer would pass to an engineer to then build it. I would try to force it [3DALL-E] to create a top view, side, front view that are somehow matching.” After generating a top view image, P18 was able to import that image as a canvas and

construct off of it. The process, where P18 sketched and extruded geometries upwards from a plane parallel to their reference image, is pictured in Fig. 13. However, we note that most generations came back angled and at perspectives unless the prompt explicitly specified viewpoints like “top view” or “flat”, and that 3DALL-E did not always capture “isometric” and “perspective” views in the technically accurate sense of those words. Nonetheless, even if generations were not drawn to perspective or as clean as technical drawings and renders usually are, participants still found inspiration in them.

Other participants used the generations more loosely. P15, liking a “3D render of a desk lamp Victorian” (Fig. 11), made the arm of their lamp skinnier as per the image. P9, observing generations from prompts such as “isometric 3d renders of a cleaning sprayer bottle” (Fig. 11), noted that they could subtract from the outer contours of their model and reduce the amount of material used, which was part of their goal to design a more sustainable spray bottle top.

7.1.2 Use Case: Textures and Renders for Editing Appearance. Participants would also edit their model appearance towards the look of generations (P13, P15, P10, P3). They could do this by applying textures within the software and dragging and dropping materials from the software’s material library onto surfaces. For example, P13 dragged wood and chrome finishes onto their speakers model and tried texturizing the faces of their model with tessellation to match the color palette and low-poly look of a generation. P5, innovatively used generations as textures to help build a 3D outdoor movie theater scene. Their scene was built out of simple geometries, and atop these geometries, they placed generations of a “jello bed” and generated portraits of pop culture characters.

P1 mentioned that 3DALL-E could be useful for product design presentations to show the function or interaction of things being designed. As P1 was making a prosthetic hand, they imported a generation and started to model atop it. Curious about how a text+image prompt would fare if it included a generation transparently covering their geometry, they generated and found compelling results.

7.1.3 Use Case: Inspiring Collaboration. Design in industry is a team effort, and while 3DALL-E was evaluated in the context of a single user, many participants acknowledged that 3DALL-E could be beneficial in teams. P16 mentioned that from their industry experience, 3DALL-E would be excellent for establishing communication between mechanical engineers and industrial designers. Mechanical engineers focus on function, while industrial designers focus on aesthetics. P16 felt that 3DALL-E could help both sides pass around design materials for discussion and establish common ground.

P13, who was an industrial designer, noted that teams could also do multi-pronged exploration with 3DALL-E. Because each team member would have individual prompting trajectories, a team could easily produce diverse searches and more variety during brainstorming. P3 mentioned that there are already points within their industry (automotives) where there are hand-offs between the people who generate design ideas and the people who execute them.

Technical sales specialist P4 also mentioned that they could instantly see 3DALL-E being useful for their clients, many of whom

have bespoke requests such as organic fixtures for restaurants and museums or optimized shapes for certain materials. While they did not see the tool as a problem solver itself, they felt it could provide ideas that would help them do the problem solving parts of their job that they love doing.

7.1.4 Use Case: Inspiring Design Considerations. 3DALL-E also inspired design considerations by making participants think about different aspects such as functionality or manufacturability. For example, P1 was looking for a wheeled robot. Seeing generations where robot bodies were varied in the number of wheels they had or how far off the ground they were made P1 think about the different amounts of motor power these robots would require. While 3DALL-E could not guarantee the feasibility of every generated design, some participants (P1, P8) liked that 3DALL-E helped them think through details such as how manufacturable a design was.

Participants also felt like they could elicit unique, out-of-the-norm designs from 3DALL-E and use it to gauge the uniqueness of their own designs. P4 wanted to design a product that did not exist in the real world yet: an ear gauge electronic for their son. They treated the model’s inability to come up with their exact vision in generations as a good thing, interpreting it to mean that the product did not exist yet and therefore had patentable value. “*We [DALL-E] started to lose a little bit when we started putting in the ‘Bluetooth ring’, which is good because that tells me...probably out there in the real world, nobody’s actually doing this...that made me feel good about the fact that I might have a predicate design in my head.*” P2, who had previously taught classes about drone design, also felt like right off the bat, 3DALL-E was able to produce unique aesthetics beyond what is typically seen in drones, something their students generally struggled to do. P15 also felt like 3DALL-E could have educational value as they looked around for ways to accomplish something they saw in a lamp generation: “*being able to reverse engineer...that is a cool learning aspect.*”

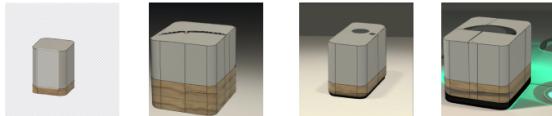
7.1.5 Weaknesses in Industry Settings. Some participants did comment that text-to-image AI may have weaknesses in applications like machining and simulation or the construction of internal components and other function-focused parts. P9 pointed out that it would be difficult to generate geometries that enclose parts, because if a user was to pass in an image prompt of that part, 3DALL-E would be unable to draw housing over it. Likewise, a participant mentioned that they could imagine 3DALL-E being used to design the facade of the car, but they did not believe that it could design a more internal component not easily describable in layman’s terms.

7.2 Incorporation into Diverse Workflows

Our exploratory study invited designers to stress test the 3DALL-E across the settings of a wide range of disciplines. Participants were impressed with the ability of the model to generate even when they passed prompts filled with technical jargon like “CNC machines”, “L-brackets”, or “drone landing gear”. Still, prompting remains very distinct from the workflows participants usually go through. Many participants described their regular design process as multiple phase progressions from low fidelity to high fidelity. They mentioned roughing out designs first, putting placeholders within robotic assemblies (P1), box blocking up to complexity (P13),

PROMPT BIBLIOGRAPHY: Design of Audio Speakers (P13-B)

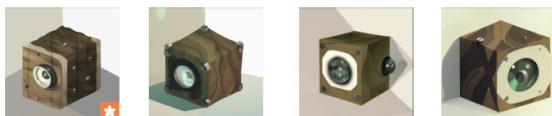
IMAGE+TEXT: Scandinavian, 3D illustration, I'm looking for a good pair of portable Bluetooth speakers



IMAGE+TEXT: Scandinavian Minimalism audio speaker with lights



IMAGE+TEXT: Scandinavian Minimalism audio speaker with lights

**VARIATION**

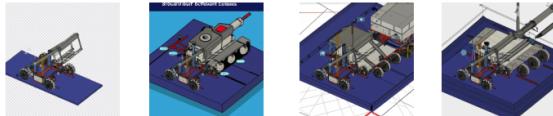
IMAGE+TEXT: Isometric Scandinavian minimalism, audio speaker with built-in light



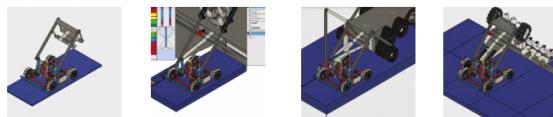
IMAGE+TEXT: Isometric, scandinavian, minimalist, audio speaker

**PROMPT BIBLIOGRAPHY:** Design of Tank Drive Robot (P1-A)

IMAGE+TEXT: Robot with tank drive system



IMAGE+TEXT: Robot with tank drive system



IMAGE+TEXT: Robot with wheels



TEXT: 3d illustration of a Roomba with wheels



TEXT: 3d illustration of a Roomba with four wheels



TEXT: 3d illustration of a Roomba with four wheels powered by motors



Figure 12: Prompt bibliographies, a design concept we propose for tracking human-AI design history. As prompts become a part of creative workflows, they may be integrated into the design histories already kept by creative authoring software. This bibliography tracks text and image prompts, as well as which generations inspired users during the tasks.



Figure 13: Snapshots of 3D design process of one participant (P18) who 3D modelled using a DALL-E generation as a reference image during T_{create} .

and redesigning from the ground up again and again (P18). 3DALL-E, in contrast, generally provided designs of the highest fidelity, short-cutting users to things they usually see later on in the design process.

7.2.1 Text Interactions in 3D Workflow. The most distinct difference in workflows is that 3DALL-E is text-focused, but text is not central to 3D design workflows, which are usually based on direct manipulation of the geometry. P13 mentioned that designers primarily operate visually. “*The only reason I really use text in an industrial design context is [for] making notations on a design...to explain what a feature is...to write a design specification...but the majority of the time is image focused.*” Because of this, P13 preferred the “image-based approach” within 3DALL-E where they could “provide it with a starting image and get variants of that”. P4, however, thought that in some respects designers *are* often engaging with text, but in the form of numbers, properties, parameters, equations, and configurations. “[We] do it in a smart way...[we] drive it with the math equation. This is something we can do in parameters, and it is very text-based.”

7.2.2 Problem Solving with 3DALL-E. P10 and P4 described their regular jobs as problem solving and finding design solutions. P10 began the study wondering if 3DALL-E could solve a problem they were facing in their job: packaging a toy truck. To do so, they like many of the participants, tried employing 3DALL-E as a problem solver. P10 tested prompts such as “*create a toy dump truck and fire truck with plastic material*” and “*protect a sphere with foam*” to see if 3DALL-E could help encase a 3D model. From the results they saw, they concluded that 3DALL-E “*was not intended to be a problem solver type of tool*”.

P13 set up image prompts as autocomplete problems. As they built an audio speaker for T_{create} , they commented that they were “*creating two pieces of geometry and using it [3DALL-E] as a connection between the two...kind of like the automated modelling command*” [34]. They also tried other innovative ways of creating image prompts, which they described as “*a hacky approach, trying to keep preserved geometries with the faces and using 3DALL-E to fill in the gaps*”. P13 also tried other complex image prompts by deleting faces and hiding parts of the geometry to create transparent places for DALL-E to fill in, even if decimating parts of the geometry would not be a natural part of their workflow.

7.2.3 Driving the Design. When AI input is added into a workflow, questions of who drives the design process and who owns the final design can arise. While P9 liked that 3DALL-E augmented their workflow with what they called dynamic feedback, they felt as though their design was being driven by the generations. “*Initially, the image did not really meet my expectation... but eventually I was also trying to not imagine anything and just depend upon what it was suggesting.*” P3 mentioned that they felt as if they were driven by DALL-E, while P15 mentioned that sometimes in the midst

of exploring, they felt they were not gravitated towards building anything.

As for ownership, many participants felt like the designs they created with 3DALL-E would still be their own. P1 stated on ownership, “*A lot of 3D modelling is stealing...borrowing premade files online, and then assembling it together into a new thing. For this robot, we borrowed these assemblies from already premade files that were sold by the company. We modelled based off of that, but the majority of this robot can be considered ours because we determined the placement.*” P13 was also not worried about ownership concerns, stating that even now, anyone can recreate any model found online, but that “*it's about the steps you go through to get there.*”

P18 mentioned that for an AI to be applied to the real world, it still takes an expert designer’s understanding of the market and customer needs. “*I would use my know-how of manufacturing processes and the market or style. My service would adopt AI as a source of inspiration rather than as the solution.*” Reflecting on if AI inspiration became mainstream, they expressed concerns that “*if everyone would converge on the same designs [because] it only learns from the input it gets from people... we might lose creativity.*”

7.3 Comparison with Existing Generative Tools

Five of 13 participants had experience with the existing generative design mode within the 3D software [35]. Generative design (GD) is an environment where the completion of a 3D design is set up like a problem: users define physical constraints and geometric filters that allow a model to be autocompleted. We did not directly compare with GD, because hardware constraints made 3DALL-E incompatible with GD. However, we did ask participants with experience in GD to compare and contrast the two.

A primary difference was that GD allows users to directly manipulate the model geometry, which differs from the text-based interaction of 3DALL-E. GD results therefore free the user from doing more modelling work. What one participant liked about GD was that “*once they set up the problem, they could just hit go... don't have to actually worry about lofting and modelling.*” However, participants mentioned that GD has a higher barrier of entry; users are burdened with calculating loads and non-conflicting constraints, which requires some understanding of physics and engineering.

“You’re [GD] focused on strength, durability of the model itself, really driven as a manufacturing task... your end result is something that’s makeable... whereas this process [3DALL-E] is more on the creative side.”

One benefit a participant (P2) mentioned of 3DALL-E was that the system allowed users to come up with outcomes far more efficiently than GD. In the span of a thirty-minute task, users were able to browse hundreds of results, with the first results coming in a matter of seconds, whereas P2 has previously had to wait multiple hours or even days for GD. P2 and P18 were enthusiastic that GD and 3DALL-E could merge. P10 suggested that one way these two tools could complement each other is if “*this tool [3DALL-E] could be used to generate shapes... pass it off to the generative design [GD] to optimize it.*”

8 DISCUSSION

Our results demonstrate enthusiasm for text-to-image tools within 3D workflows. With 3DALL-E, people had a tool to help them combat design fixation and get a variety of use cases and inspiration. Furthermore, we elaborated different patterns of prompting that can identify endpoints at which text-to-image can help. In measuring prompt complexity, we showed that many prompts fall within a range of two to six concepts. We were also able to capture participant intentions and rich design history in the form of prompt bibliographies, as shown in Fig. 12. The following discussion focuses on best practices for helping users bring their own work into AI-assisted design workflows and the implications of these workflows.

8.1 Prompt Bibliographies

A strength of studying 3D workflows was that there was no conflict between the AI and human on the canvas, as the AI had no part in the physical realization of the design. We believe this helps mitigate ownership concerns and make text-to-image AI very promising for 3D design tools.

However, ownership over generations from text-to-image AI can still be a gray area in some respects. Currently, there is no way to tell how heavily an AI-generated image borrows from existing materials. It is difficult to embed signatures within AI-generated images. DALL-E generations come with small watermarks at the bottom of each image, though these can be easily cropped out. As AI-generated art becomes more mainstream and prevalent on platforms, it is important to develop practices of data provenance [20]. Even if right now we cannot understand what directly inspires the contents of an AI generation, we propose the notion of *prompt bibliographies* to provide information on what informed our designs and separate out which contributions were human and which were AI. These can work to clarify ownership and intellectual property concerns.

Prompt bibliographies could likewise help enrich the design histories that software tools provide, which generally capture commands and actions—but not intentions. The bibliographies can be merged within the history timeline features that are present in tools like Fusion 360 and Photoshop, helping prompting integrate better with the traditional workspaces of creative tools.

Sharing prompt bibliographies with their outcomes (i.e. 3D models) can also help respect all the parties that are behind these AI systems. During pretraining, these models were built off the backs of human art and images; end users can easily query for the styles of artists (as they already do) and create derivative works that dilute the pool of images attributed to artists. For now, potentially the best methods of watermarking AI tools can come about from due diligence and effort on the part of the creators [38].

8.2 Enriching creative workflows with text

The advancements in prompting may push text prompting as a type of interaction into creative tools, even if creative workflows have traditionally not revolved around text. In 3DALL-E, we show the benefit of having a language model scaffold the prompting process. By giving the user fast ways to query and gesture towards what

an AI is most likely to understand (as 3DALL-E did with the highlighted text options), we enable users to have more opportunities to understand what language may work best with an AI. At the same time, 3DALL-E helped users easily reach the design language of their domain, be it robotics or furniture design. In the quantitative survey results, participants felt it was easy to come up with prompts near unanimously for T_{edit} and unanimously for T_{create} .

It is also important to scope out places in the workflow where AI may be of most assistance, whether it be through use cases or through an understanding of what points of entry in a workflow make the most sense. Our survey results reflect that the system produced a slightly more positive experience when it was introduced earlier on in the process. This was corroborated by many participants who said they saw this tool being most helpful in the early stages of design. Ill-placed AI assistance, such as suggesting divergent directions when a model is nearly complete or providing incoherent results that do not ideally respond to prompts could be unproductive for designers. Well-placed AI assistance, however—such as early stage ideation with GPT-3, trying a text-only prompt to pivot directions, or carefully setting up an image prompt for DALL-E to fill in—can be greatly constructive and address pain-points like design fixation that designers feel today. Furthermore, if we understand the scope of the tasks we want AI to handle within a workflow, such as having GPT-3 suggest different parts of a model or having DALL-E generate reference images from front, side, and top views, we can better fit general purpose models to their task. This is not to imply that we need to finetune the models, but we can have stronger checks on the prompt inputs and generation outputs if we understand what is within scope of the task. For example, when P16 wanted a “flat head” screwdriver, they were returned results about a medical syndrome—something that could be avoided with content filtering guards checking for relevance to 3D design. In this way, AI models do not have to bear the full burden of providing good and ethical answers, and we can have multiple checkpoints for propriety.

8.3 Generalizability

The design workflow posed in 3DALL-E is generalizable and can easily be used as a blueprint for text-to-image AI integration with different design software. The idea behind surfacing 3D keywords from software related data also introduces ideas for how prompts can be tailored towards the technical vocabulary of a software. The idea of passing in image prompts is also easily extendable to different creative tools, even those outside of the 3D space. For example, graphic editing tools can pass in image prompts based on active layers chosen by a user. Animation software and video editors can send in choice frames for anchored animations and video stylization. A takeaway of this paper is to take advantage of the complex hierarchies that users build up as they design, such as the way 3DALL-E takes advantage of the fact that 3D models are generally assemblies of parts. With 3DALL-E, users could isolate parts and send clean image prompts without the burden of erasing or masking anything themselves.

We demonstrated the efficacy of this tool at supporting a diverse set of potential end users: mechanical engineers, industrial designers, roboticists, machining specialists, and hobbyist makers.

3DALL-E’s interdisciplinary nature is both a strength of AI pretraining as well as the ability of designers to make integrative leaps to meet the AI halfway [62]. Software companies and their customer clients may each have their own domains and internal tools, but the modular nature of 3DALL-E facilitates its integration across different settings. Additionally, the modular nature of 3DALL-E in Fusion 360 demonstrates an idea of separating out AI assistance from traditional non-AI direct manipulation features. Lastly, the text-based nature of the tool and its ready acceptance with designers demonstrates how text interactions can facilitate low threshold, high ceiling design tools [48].

8.4 Future Work and Limitations

In terms of limitations, there were times during the study when we experienced technical difficulties. For example, some participants had their DALL-E results cancelled. Moreover, when participants tried to compare the generative design environment with 3DALL-E, the software crashed, so we were unable to directly compare 3DALL-E with GD. However, the existence of GD, a cloud-based generative design tool, shows that there are already users within industry who willingly navigate trade-offs of exploration and control. As such, future work can explore how these tools can merge, as 3DALL-E has the potential to help with text-based exploration of GD outcomes.

Data privacy will also be a key concern in the future. Design know-how and details are the intellectual property of companies and is safeguarded by high-value product industries (i.e. cars). We asked participants to use non-sensitive files, but in the future it will become important to understand how intellectual property passed to AI systems can be protected and not given as free training data. While each prompt needs to be examined for content policy and ethics adherence, there are looming trade-offs to be made in data privacy and AI regulation.

The work primarily investigated how well text-to-image tools could be embedded in a workflow where the designs grow increasingly complex. Future work can investigate whether text-to-image AI can provide inspiration that can hold across many different image frames. Rather than producing one image frame of inspiration, we can look at how these models can be embedded into animation pipelines, places where creators are already partially used to automation and computational aids.

9 CONCLUSION

3DALL-E introduced text-to-image AI into 3D workflows and was evaluated in an exploratory study with 13 designers. This study elaborated a number of use cases for text-to-image AI from providing reference images to facilitating collaboration to inspiring design considerations. From participant prompts, we observed different types of prompting patterns depending on whether the user engaged with 3DALL-E first, last, or throughout their process. Furthermore, we provided measures of prompt complexity across participants and propose a concept for tracking human-AI design history through prompt bibliographies.

ACKNOWLEDGMENTS

We thank the AI Lab at Autodesk Research and OpenAI for their help.

REFERENCES

- [1] Adverb. 2021. Advadnoun. <https://twitter.com/advadnoun>
- [2] Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fournier, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul Bennett. 2019. Guidelines for Human-AI Interaction. 1–13. <https://doi.org/10.1145/3290605.3300233>
- [3] Autodesk. 2022. Autodesk Screencast. <https://knowledge.autodesk.com/community/screencast> Retrieved September 15, 2022.
- [4] Gwern Branwen. 2020. Gpt-3 creative fiction. <https://www.gwern.net/GPT-3>
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. <https://doi.org/10.48550/ARXIV.2005.14165>
- [6] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Séan Ó hEigearthaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shaguun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. <https://doi.org/10.48550/ARXIV.2004.07213>
- [7] Angel X. Chang, Mihail Eric, Manolis Savva, and Christopher D. Manning. [n. d.]. SceneSeer: 3D Scene Design with Natural Language. <https://doi.org/10.48550/ARXIV.1703.0050>
- [8] Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. 2013. Attribit: Content Creation with Semantic Attributes. *ACM Symposium on User Interface Software and Technology (UIST)* (Oct. 2013).
- [9] Siddhartha Chaudhuri and Vladlen Koltun. 2010. Data-Driven Suggestions for Creativity Support in 3D Modeling. *ACM Trans. Graph.* 29, 6, Article 183 (dec 2010), 10 pages. <https://doi.org/10.1145/1882261.1866205>
- [10] Erin Cherry and Celine Latulipe. 2009. The creativity support index. 4009–4014. <https://doi.org/10.1145/1520340.1520609>
- [11] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. 2001. Using Information Scent to Model User Information Needs and Actions and the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) (CHI '01). Association for Computing Machinery, New York, NY, USA, 490–497. <https://doi.org/10.1145/365024.365325>
- [12] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. <https://doi.org/10.48550/ARXIV.2202.04053>
- [13] Bob Coyne and Richard Sproat. 2022. WordsEye: an automatic text-to-scene conversion system. <https://doi.org/10.1145/383259.383316>
- [14] Katherine Crowson. 2021. afiaka87/clip-guided-diffusion: A CLI tool/python module for generating images from text using guided diffusion and CLIP from OpenAI. <https://github.com/afiaka87/clip-guided-diffusion>
- [15] Katherine Crowson. 2021. Rivers Have Wings. <https://twitter.com/RiversHaveWings>
- [16] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. *arXiv preprint arXiv:2204.08583* (2022).
- [17] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. In *CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3491102.3502141>
- [18] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khắc, Luke Melas, and Ritobrata Ghosh. 2021. DALLE Mini. <https://doi.org/10.5281/zenodo.1234>
- [19] Manoj Deshpande, Eric Sauda, and Mary Lou Maher. 2020. Towards Co-Build: An Architecture Machine for Co-Creative Form-Making.
- [20] Daniel Deutch, Tanu Malik, and Adriane Chapman. 2022. Theory and Practice of Provenance. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) (SIGMOD '22). Association for Computing Machinery, New York, NY, USA, 2544–2545. <https://doi.org/10.1145/3514221.3524073>
- [21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* [cs.CL]
- [22] Alaa El-Nouby, Shikhar Sharma, Hannes Schulz, R Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Y. Bengio, and Graham Taylor. 2018. Keep Drawing It: Iterative language-based image generation and editing.
- [23] Patrick Esser, Robin Rombach, and Björn Ommer. [n. d.]. Taming Transformers for High-Resolution Image Synthesis. <https://doi.org/10.48550/ARXIV.2012.09841>
- [24] Thomas Funkhouser, Michael Kazhdan, Philip Shilane, Patrick Min, William Kiefer, Ayellet Tal, Szymon Rusinkiewicz, and David Dobkin. 2004. Modeling by Example. *ACM Trans. Graph.* 23, 3 (aug 2004), 652–663. <https://doi.org/10.1145/1015706.1015775>
- [25] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. [n. d.]. Greater Creative Control for AI image generation. <https://ai.facebook.com/blog/greater-creative-control-for-ai-image-generation/>
- [26] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) (UIST '15). Association for Computing Machinery, New York, NY, USA, 489–500. <https://doi.org/10.1145/2807442.2807478>
- [27] Songwei Ge and Devi Parikh. 2021. Visual Conceptual Blending with Large-scale Language and Vision Models. *arXiv:2106.14127* [cs.CL]
- [28] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. *Mental Models of AI Agents in a Cooperative Game Setting*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376316>
- [29] Katy Ilonka Gero, Vivian Liu, and Lydia B. Chilton. 2021. Sparks: Inspiration for Science Writing using Language Models. <https://doi.org/10.48550/ARXIV.2110.07640>
- [30] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. 2019. Interactive Sketch and Fill: Multiclass Sketch-to-Image Translation. *arXiv:1909.11081* [cs.CV]
- [31] Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2010. Chronicle: Capture, Exploration, and Playback of Document Workflow Histories. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) (UIST '10). Association for Computing Machinery, New York, NY, USA, 143–152. <https://doi.org/10.1145/1866029.1866054>
- [32] S. G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in psychology* 52 (1988), 139–183.
- [33] Autodesk Inc. [n. d.]. Fusion 360. <https://help.autodesk.com/view/fusion360/ENU/?guid=GUID-7B5A90C8-E94C-48DA-B16B-430729B734DC>
- [34] Autodesk Inc. 2022. Autodesk Fusion 360 faster performance and quality of life updates. <https://www.autodesk.com/products/fusion-360/blog/usability-and-performance-improvements-fusion-360/>
- [35] Autodesk Inc. 2022. Generative design for manufacturing with Fusion 360. <https://www.autodesk.com/solutions/generative-design/manufacturing>
- [36] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. *FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411764.3445093>
- [37] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J. Cai. 2022. PromptMaker: Prompt-based Prototyping with Large Language Models. <https://doi.org/10.1145/3491101.3503564>
- [38] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. <https://doi.org/10.48550/ARXIV.2006.06676>
- [39] Tero Karras, Samuli Laine, and Timo Aila. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. <https://doi.org/10.48550/ARXIV.1812.04948>
- [40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958* [cs.CV]
- [41] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3491102.3502030>
- [42] Vivian Liu and Lydia B. Chilton. 2021. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. *arXiv:2109.06977* [cs.HC]
- [43] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. <https://doi.org/10.48550/ARXIV.2204.09007>
- [44] Maria Teresa Llano, Mark d'Inverno, Matthew Yee-King, Jon McCormack, Alon Ilzar, Alison Pease, and Simon Colton. 2022. Explainable Computational Creativity. (2022). <https://doi.org/10.48550/ARXIV.2205.05682>
- [45] Ryan Louie, Any Cohen, Cheng-Zhi Anna Huang, Michael Terry, and Carrie J. Cai. 2020. Cococo: AI-Steering Tools for Music Novices Co-Creating with Generative Models. In *HAI-GEN+user2agent@UII*.
- [46] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. <https://doi.org/10.48550/ARXIV.2111.09734>

- [47] Ryan Murdock. [n. d.]. lucidrains/big-sleep: A simple command line tool for text to image generation, using OpenAI's CLIP and a BigGAN. Technique was originally created by <https://twitter.com/advadnoun>. <https://github.com/lucidrains/big-sleep>
- [48] Brad Myers, Scott E. Hudson, and Randy Pausch. 2000. Past, Present, and Future of User Interface Software Tools. *ACM Trans. Comput.-Hum. Interact.* 7, 1 (mar 2000), 3–28. <https://doi.org/10.1145/344949.344959>
- [49] OpenAI. 2022. Dall-E: Introducing outpainting. <https://openai.com/blog/dall-e-introducing-outpainting/>
- [50] Guy Parsons. 2022. The DALL-E 2 prompt book. <https://dallery.gallery/the-dalle-2-prompt-book/>
- [51] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106 (10 1999), 643–675. <https://doi.org/10.1037/0033-295X.106.4.643>
- [52] Han Qiao, Vivian Liu, and Lydia Chilton. 2022. Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art. In *Creativity and Cognition* (Venice, Italy) (C&C '22). Association for Computing Machinery, New York, NY, USA, 15–28. <https://doi.org/10.1145/3527927.3532792>
- [53] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Learn, Imagine and Create: Text-to-Image Generation from Prior Knowledge. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2019/file/d18f655c3fce66ca401d5f38b48c89af-Paper.pdf>
- [54] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning Text-to-image Generation by Redescription. <https://doi.org/10.48550/ARXIV.1903.05854>
- [55] 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. [arXiv:2103.00020 \[cs.CV\]](https://arxiv.org/abs/2103.00020)
- [56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. <https://doi.org/10.48550/ARXIV.2204.06125>
- [57] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. [arXiv:2102.07350 \[cs.CL\]](https://arxiv.org/abs/2102.07350)
- [58] 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 (CVPR)*. 10684–10695.
- [59] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. <https://doi.org/10.48550/ARXIV.2205.11487>
- [60] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Karnal Rahimi Malekshan. 2021. CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation. <https://doi.org/10.48550/ARXIV.2110.02624>
- [61] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. 2018. ChatPainter: Improving Text to Image Generation using Dialogue. <https://doi.org/10.48550/ARXIV.1802.08216>
- [62] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* (jan 2022). [https://doi.org/10.1145/3511599 Just Accepted](https://doi.org/10.1145/3511599)
- [63] Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 582, 11 pages. <https://doi.org/10.1145/3411764.3445219>
- [64] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. <https://doi.org/10.48550/ARXIV.2203.06566>
- [65] Tongshuang Wu, Michael Terry, and Carrie J Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. <https://doi.org/10.1145/3491102.3517582>
- [66] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2020. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation. <https://doi.org/10.48550/ARXIV.2012.03308>
- [67] Kai Xu, Hanlin Zheng, Hao Zhang, Daniel Cohen-Or, Ligang Liu, and Yueshan Xiong. 2011. Photo-inspired model-driven 3D object modeling. In *ACM SIGGRAPH 2011 papers on - SIGGRAPH '11*. ACM Press, Vancouver, British Columbia, Canada, 1. <https://doi.org/10.1145/1964921.1964975>
- [68] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2017. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. <https://doi.org/10.48550/ARXIV.1711.10485>