Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art

Han Qiao
hqiao29@gmail.edu
Columbia University
New York, New York, USA

Vivian Liu
vivian@cs.columbia.edu
Columbia University
New York, New York, USA

Lydia B. Chilton
chilton@cs.columbia.edu
Columbia University
New York, New York, USA

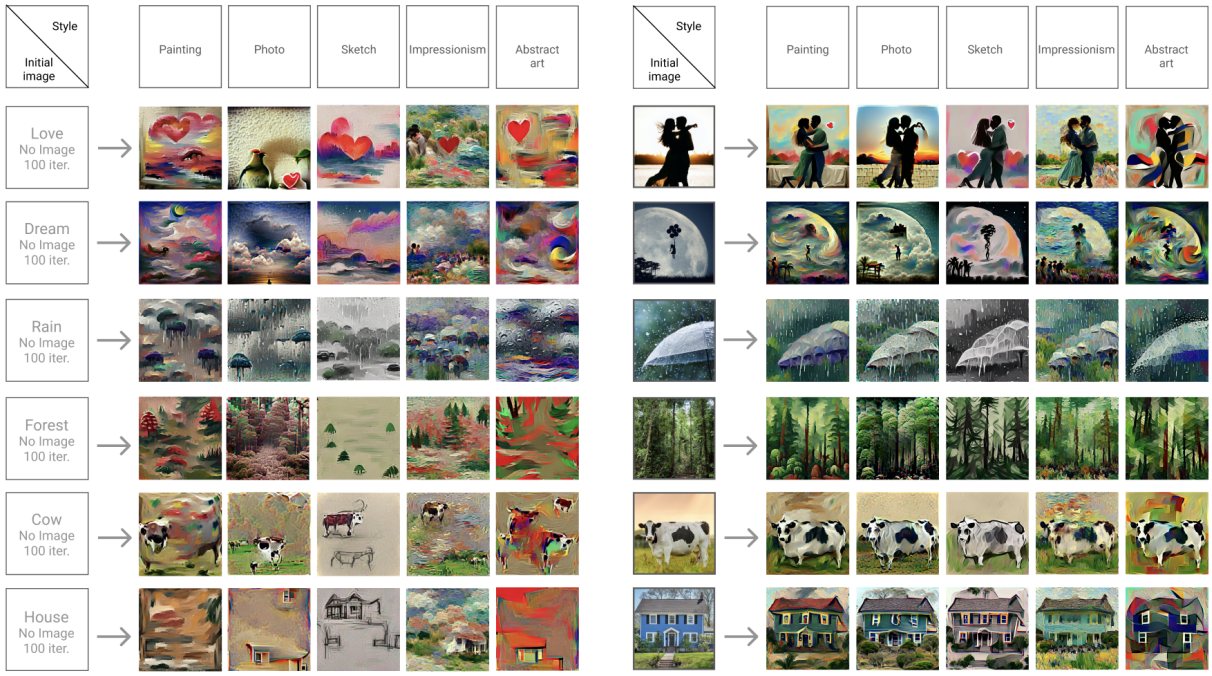

Figure 1: AI generated art with and without initial images for three types of subjects: abstract (love, dream), concrete plural (rain, forest), and concrete singular (cow, house). Each row contains one subject rendered in five art styles (painting, photo, sketch, impressionism, abstract art). An annotation study found that initial images significantly improved subject representation across all subject types with concrete singular subjects showing the most improvement.

ABSTRACT

Advances in text-to-image generative models have made it easier for people to create art by just prompting models with text. However, creating through text leaves users with limited control over the final composition or the way the subject is represented. A potential solution is to use image prompts alongside text prompts to condition the model. To better understand how and when image prompts can improve subject representation in generations, we conduct an annotation experiment to quantify their effect on generations of abstract, concrete plural, and concrete singular subjects. We find that initial images improved subject representation across all subject types, with the most noticeable improvement in concrete singular subjects. In an analysis of different types of initial images, we find that icons and photos produced high quality generations of different aesthetics. We conclude with design guidelines for how initial images can improve subject representation in AI art.

C&C '22, June 20–23, 2022, Venice, Italy

ACM ISBN 978-1-4503-9327-0/22/06…$15.00
https://doi.org/10.1145/3527927.3532792

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Neural networks*; • **Applied computing** → **Media arts**.

KEYWORDS

text-to-image, multimodal generative models, computational creativity, prompt engineering, design guidelines, AI co-creation

ACM Reference Format:
Han Qiao, Vivian Liu, and Lydia B. Chilton. 2022. Initial Images: Using Image Prompts to Improve Subject Representation in Multimodal AI Generated Art. In *Creativity and Cognition (C&C '22), June 20–23, 2022, Venice, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3527927.3532792

1 INTRODUCTION

The development of text-to-image generative models has expanded the possibilities to create intriguing artworks with the help of AI. The DALL-E model [12] introduced by OpenAI in 2021 demonstrated how high quality images can be declaratively generated using simple text and image prompts. An emerging design area at the intersection of AI and art has since evolved on Twitter, Reddit, and Github. Online, practitioners share open source models, generations, and suggestions for prompts and model parameterizations. Recently, researchers have also systematically delved into prompt engineering and parameter tuning to better understand the model and create design guidelines. Important insights, including certain text prompt templates such as "<Subject> in the style of <Style>," have been shown to produce high quality generations. (See Figure 1 for examples.) [22]

Although following design guidelines for text prompts can improve AI generation, the user has no control over the composition of the image or the way the subject is represented. There are unique failure modes that come with prompting generation using only text. Text-to-image generations often suffer from poor representation of the subject and the fact that text can often be misinterpreted into irrelevant generations when there are multiple meanings for prompt words. Furthermore, because these AI generations come from large scale models that span thousands of classes of data, there is no way to guarantee the correct natural structure of an image. A generation of a dog could emerge with the head, body, and legs all signifying a dog, but in a deconstructed disarray. The automatic nature of generation is also contrary to how most visual artwork has been made throughout human history–through the manual specification of what detail goes where. When people design graphics or paint canvases, the first things that come to their mind are not descriptive texts of their design but usually sketches or spatial layouts.

To address the above problems of misinterpretation and poor composition in generations, we examine a mode of control that has to our knowledge not yet been studied: image prompts for text-to-image models. While significant work has been done within natural language processing around the prompt engineering of text prompts for text generation and while [22] looked at the prompt engineering of text prompts for image generation, we study how text and image prompts can be concurrently used for image generation. The image prompts are also referred to as "initial images" — an image given to the model that initializes the generation from the chosen image instead of random noise. As co-creative systems are now complex enough to handle multimodal inputs for prompts, it is important for us to understand what each mode of input is best equipped at handling and how to best support user control for future systems.

Text and image express different elements of creativity. In text, we can summarize the intention of a creative goal and in image we dictate the realization of that goal in compositions. [22] expressed guidelines for how multimodal generative frameworks like VQGAN+CLIP can be guided by intentions through prompts. In this paper, we try to understand to what degree we can control the final generation with initial images, steering the composition and realization of a user's creative goals. We contribute work that helps understand when it is better to guide multimodal co-creative systems with image and when it is better to guide a multimodal system with text. Specifically, in this paper, we conduct an annotation study looking at subjects of different levels of abstractness and initial images of different levels of detail. To systematically analyze how the model performs on a wide variety of subjects, we categorize the subjects into 3 groups: abstract, concrete singular, and concrete plural. These groupings were based on cognitive science theories around how we cognitively engage with visual art works. Kandel writes in *Reductionism in Art and Science* [18] that abstract and figurative artworks are processed differently by the human brain. Kandel states that figurative artworks invoke bottom-up processing, enabling us to recognize elements from the physical world in artwork. Abstract artworks invoke top-down processing, inviting us to engage our associative memory and imagination to make sense of what we have seen. Because images with abstract or concrete subjects are so distinct in the ways they cognitively engage us, we expect that creating them may require different strategies of interaction. In the case of text-to-image models, this interaction occurs in the form of text and image prompts. To systematically analyze how the model performs with different types of image prompts, we look in particular at icons (minimal and symbolic images) and photos (detailed and realistic images).

In our annotation study, we use text prompts of the form: "<Subject> in the style of <Style>" (i.e. "cow in the style of abstract art.") and look at whether the type of initial images have a significant effect on the quality of subject representation in generations across different subject types. Using icons and photos of subjects, we test 6 abstract subjects (i.e. "love" and "dream"), 6 concrete singular subjects (i.e. "dog" and "car"), and 6 concrete plural subjects (i.e. "forest" and "fire"). We have two annotators with art expertise rate the generations across the conditions.

From the annotation results, we contribute the following contributions:

- Quantitative analysis to find cases when initial images significantly improve subject representation. More specifically, we found: 1) Initial images significantly improved the quality of subject representation across all subject types. 2) For concrete singular subjects, initial images contributed the greatest improvements. 3) For abstract subjects and abstract art styles, generations with icon initial images received significantly higher ratings than photo initial images.
- Qualitative analysis of the aesthetic differences, benefits and drawbacks of pairing certain types of initial images across different types of subjects and styles.
- Design guidelines addressing how users can use initial images to produce better generations given the nature of the subject and the initial image

We conclude with a discussion about how these results are relevant for understanding user control of multimodal models and expanding conversation about prompt engineering into images.

2 RELATED WORK

2.1 Cognition and Visual Perception of Artwork

A significant body of work within cognitive science exists at the intersection of cognition, image processing, and aesthetics addressing how we cognitively engage with visual media. It has been theorized in [6] that our brains originally evolved routes for image processing as a survival mechanism, so we could have a constantly updating mental representation of the world. However, eventually as culture began to produce artistic images, we developed cognitive markers for what is a natural image and what is an artistic image. Artistic style is one such cognitive marker; another important finding that has been reinforced by multiple researchers in [11] and [5] is that the way we cognitively process artistic style is distinct from the way we process the subject matter of work.

Further work examining artistic styles along the abstract representational split has also found that abstract art is processed differently compared to representational art. A seminal neuroaesthetics study done by Kawabata and Zeki [21] studied different types of representational paintings (landscape, portrait, and still life) and found that they activated different localized and category-specific parts of our brain. In contrast, abstract art could not be attributed to any one region or unique brain activity, a result reinforced by [36]. Researchers studying behavior through eye-tracking have found further differences between abstract and representational art. We tend to engage with representational art that converges on the salient features (i.e. a nose or a tree) [24], while abstract art tends to invite a more uniform gaze. [35] Because we process abstract and representational work differently, we need to scaffold generative processes accordingly given which type of image is our objective.

2.2 Generative Design and AI Generated Art

Generative design presented a new form of design as computation became increasingly embedded in artistic workflows. Generative designs are designs that are produced in part or in whole by code. The design process for a user often involves the exploration of a design space and the choice of one or many proposed design solutions. [3] summarized generative design into two categories: algorithmic and neural. The first approaches of generative design were algorithmic, in that artists specified constraints for the computer and the computer created parameterized designs. Early pioneers of this computer art form included artists active in the such as George Nees, Frieder Nake, and Vera Molnar. In the 1950's, many of these pioneers used plotters, drawing machines, and computers to create systematically determined visual artwork which often employed randomness and arrived at geometric abstract styles. The terminology to describe generative design has evolved over time from computer art to computer graphics, algorithmic art, and interactive art [28].

Algorithmic art has been studied within human-computer interaction under the creativity support context through systems such as Dreamlens [25], Design Adjectives [34], and neurosymbolic system in [3]. Dreamlens helped users explore a generative design space of thousands of 3D computer aided designs for tables, guiding users through the high volume of designs using parameterized data visualization methods. [34] implemented a system to design materials, fonts, and animated backgrounds by leveraging user-provided examples of preferences and bayesian inference. [3] provided a classic example of how an algorithmic art system can generate an exponential amount of geometric abstract designs, which they used to provide data for machine learning.

The other class of generative design is neural generative design, which is largely data driven. Early creativity support systems in this vein included Attribit, a statistically driven animal shape generator which mixed and matched animal parts into creatures.[8] Neural generative design has since advanced rapidly in tandem with machine learning. For example, in 2013 Google released DeepDream, an activation maximization technique made for visualizing neural networks that happened to produce pareidolic images of visual interest for artists.[26] The popularization of style transfer [14] methods soon after provided another method of artistic interaction, allowing users to filter images with specific artist techniques and color palettes.

The advent of generative adversarial networks (GANs) [17] established a new period in neural generative design, as artists leveraged the continuous nature of the embedding space in GANs such as [1, 7, 20] to create interactive and animated visual media. [4] Research prototypes such as [16] let users interactively co-create with GANs through sketch, giving users the ability of direct manipulation. Another research system, crea.blender, is an example of a novel co-creative game that allowed players to "blend" existing images into new images using BigGAN, allowing researchers to better understand how constraints can be incorporated in co-creative systems. [33] A special edition of the game, crea.blender SDG, incorporated StyleGAN2, allowing users to blend landscape images in a goal-oriented fashion for sustainability.

Suites of systems built on top of GANs have helped researchers elaborate computational creativity frameworks. A past study looking casual creation looked at the landscape of systems and identified a number of interaction techniques such as one-touch creativity, vague creation and mutant shopping. [9, 31] Other literature supplementing system work has focused on how to evaluate generative artifacts [19] and how to build more usable ones, by proposing design principles with mental models, long-term memory, and role-taking involved [23].

2.3 Text-to-Image Generation

The latest advancements in machine learning have centered around the integration of natural language intelligence and visual intelligence into singular ML frameworks. CLIP was an example of a multimodal network which learned an embedding space for both text and images using a contrastive loss. [32] DALL-E, a network utilizing CLIP, illustrated how prompts could be interpreted into images by using the embedding space as an evaluator during optimization. [2] released the first open-source text-to-image model, by combining BigGAN with CLIP, after which came a slew of other generative frameworks that tried to achieve the same purpose: VQGAN+CLIP, DeepDaze, BigSleep, and CLIP diffusion models. [10, 27, 29]These models pair different generative models with CLIP. For example, StyleCLIP allows users to conduct text-driven manipulation of StyleGAN imagery [30] , while VQGAN+CLIP [29]

allows users to generate from a vision transformer that learned a codebook of visual concepts. Recent research projects have also expanded into optimization-based synthesis methods that do not require prior training of a generative model – CLIPDraw utilizes a set of RGBA Bézier curves to create drawings that visually represent text prompts, producing a completely different visual aesthetic compared to the other ML frameworks [13].

In some capacity, text-to-image systems can be utilized like style transfer, especially when text prompts include style keywords and initial images are provided. Style transfer methods are often excellent at translating style information from a source image to a target, while retaining subject detail. However, text-to-image systems are distinct from style transfer as a framework to understand, because they involve text as a method of interaction as well, making the user experience multimodal.

Given the novel nature of the technology, few works have studied human-computer interaction principles which can support the usability of text-to-image systems. One work [22] conducted a series of experiments to elaborate design guidelines for prompt engineering and parameter setting text-to-image models. These design guidelines validated that certain families of text prompts (i.e. <Subject> in the style of <Style>) can be used to control the content and aesthetic of the image. They further found that the abstract or concrete nature of certain prompt keywords (such as the <Style> and <Subject>) can significantly interact and influence the quality of the generations. Another work by Ge and Parikh [15] examined the usability of text-to-image systems in creating visual blends. They found that text prompts alone could successfully blend different visual concepts together into one image, when they used prompts embedding shape information. Prompt engineering and parameter tuning present many open questions for human-computer interaction researchers as a form of interaction emerging in dominance.

3 EXPERIMENT METHODOLOGY

3.1 Research Questions

Because we engage with visual depictions of abstract and concrete subjects differently, we wanted to understand what the effect of using a initial image is across abstract subjects, concrete plural subjects, and concrete singular subjects. Abstract subjects involve concepts that lack physical references such as sadness, dream and thought. Concrete singular subjects involve things that have physical references, such as cars, houses, and dogs. Concrete plural subjects also have physical references, but their forms do not capture a single unit, often picturing uncountable concepts such as a forest, an ocean, or a fire.

In our experiment, we addressed the following research questions:

RQ 1: Do initial images improve subject representation for different subject types?

- **H1.1** For abstract subjects, initial images would not significantly improve subject representation in generations.
- **H1.2**: For concrete plural subjects, initial images would not significantly improve subject representation in the generations.
- **H1.3**: For concrete singular subjects, initial images would improve subject representation in generations.

RQ 2: What types of initial images, icon or photo, will lead to better subject representation across different types of subjects?

- **H2.1**: For abstract subjects, photo initial images would significantly improve subject representation in generations more than icon initial images.
- **H2.2**: For concrete plural subjects, photo initial images would significantly improve subject representation in generations more than icon initial images.
- **H2.3**: For concrete singular subjects, there will be no significant difference in subject representation between generations from icon initial images and generations from photo initial images.

3.2 Methodology

To address our research questions, we generated 720 images from a configuration of VQGAN+CLIP pretrained on Imagenet with the 16384 sized codebook [29]. All generations created from text prompts in the form of "<Subject> in the style of <Style>." We selected five art styles that spanned a variety of aesthetics and art forms: painting, photo, sketch, impressionism, and abstract art. We looked at 18 subjects, with equal representation across the abstract, concrete singular, and concrete plural subject categories. Abstract subjects included love, happiness, sadness, dream, hate and thought. Concrete plural subjects included forest, rain, mountain, fire, road, and ocean. Concrete singular subjects included car, dog, house, cat, flower, and cow.

For each combination of <Subject> and <Style>, we generated a set of eight images to test the effect of initial images. Four generations were generated without initial images, from crossing two different random initializations with two different lengths of optimization (100 iterations and 300 iterations) [1]. Four other generations were generated from crossing two different random initializations with two different initial images: icons or photos. We looked at icons and photos as different types of initial images, because they captured different levels of detail and stylization. Icons provide minimal details but are more stylized, while photos provide maximal detail with no stylization. For consistency across icons, we chose icons from the Noun Project, one of the most popular vector icon repositories online, by searching for the subject and picking two of the top results. Likewise for photos, we searched Google Images for the subject and picked two of the top results as our photo initial images. All four generations with initial image prompts were optimized for 100 iterations. To prevent any effect of ordering, we randomized the order of the generations within each set and shuffled the sets before presentation to annotators. Two annotators with experience in art and design were asked to rate generations with the following question: "on a scale of 1-5, how well does the image

[1]While the number of iterations does affect the generation, we chose these numbers deliberately based on empirical work done in [22]. In this prior work, it is expressed that from a range of 100 to 1000 iterations, the generations are most highly rated in terms of user preference when they are optimized for a shorter range of iterations–(100-500 iterations). The authors stated that more optimization beyond that range had greater potential to lead to artifacts and intensified contrasts that can make generations less aesthetically pleasing. For this reason, we chose short fixed iteration lengths of 100 and 300. Specifically, we used 100 and 300 iterations for the no initial image condition and 100 for the condition with initial images

represent the subject?". They answered this question according to the rubric shown in Table 1. Annotator 1 has more than 10 years of art experience specifically in fine art. Annotator 2 has more than 5 years of art experience in digital and new media art. All annotators were compensated $20/hour for the length of time it took them to complete the task.

3.3 Results

3.3.1 *RQ 1 Quantitative Analysis: Do initial images improve subject representation?*

To make sense of the agreement in our annotations before jumping into quantitative analysis, we used a linearly weighted Cohen's Kappa and measured interrater reliability. The weighted Cohen's Kappa for the two annotators' ratings was 0.47, which indicated moderate agreement. Based on this result, we found it valid to average the ratings from the two annotators in the proceeding analysis.

To answer *RQ1: Does using initial images improve subject representation within generations across different subject types*, we used Mann-Whitney U tests to compare the ratings of generations with initial images and the ratings of generations without initial images. To test H1.1, we used a Mann-Whitney U test to compare the distribution of mean ratings of generations with initial images and generations without initial images in the abstract subject category. We found a significant difference (p-value < 0.001), with a mean rating of 2.73 for generations with initial images and a mean rating of 1.65 for generations without initial images. The result differs from our original hypothesis. To test H1.2, we again used a Mann-Whitney U test to compare the mean ratings of generations with initial images and generations without initial images in the concrete plural subject category. We found a significant difference between the two groups (p-value < 0.001), with a mean rating of 3.58 for generations with initial images and a mean rating of 2.79 for generations without initial images. The result again differs from our original hypothesis H1.2. Lastly, we conducted a Mann-Whitney U test for H1.3, comparing the mean ratings of generations with and without initial images in the concrete singular category. The difference in mean ratings is again significant, with 3.64 for generations with initial images and 1.85 for generations without initial images. The result aligns with our original hypothesis. It is noteworthy that although our result indicates H1.1 and H1.2 were wrong, the result is still intuitive because the differences in mean ratings were much lower for the abstract and concrete plural subject categories than that of the concrete singular subject categories. The mean ratings are shown in Table 2 and Figure 2.

3.3.2 *RQ 1 Qualitative Analysis: Do initial images improve subject representation?*

Concrete singular subjects showed the greatest improvement in mean rating when initial images were used among all three subject categories. When concrete singular subjects (like "cow" and "flower") were generated without initial images, they received lower ratings when they represented the subject in a fractured and incoherent manner, a common failure mode expressed in [22] (see Figure 3). While the features of the subject could be expressed, they were often composed in incorrect or unnatural arrangements. For example, in the first two rows of the bottom group in Figure 3, we see that in generations of the "cow" subject made without a initial image, salient features of cows would appear, but in disarray and not with the natural structure of any cow. On the other hand, if we look at the next two rows in Figure 3, studying generations made with icon and photo initial images, we see key features of cows presented with an identifiable subject. We also see this pattern appearing in the other concrete singular subjects such as cat, dog, car, and so on. We illustrate another example for the concrete singular subject of "flower" in Figure 3, demonstrating how initial images improved subject representation even when the difference in mean rating between initial and no initial images was less pronounced.

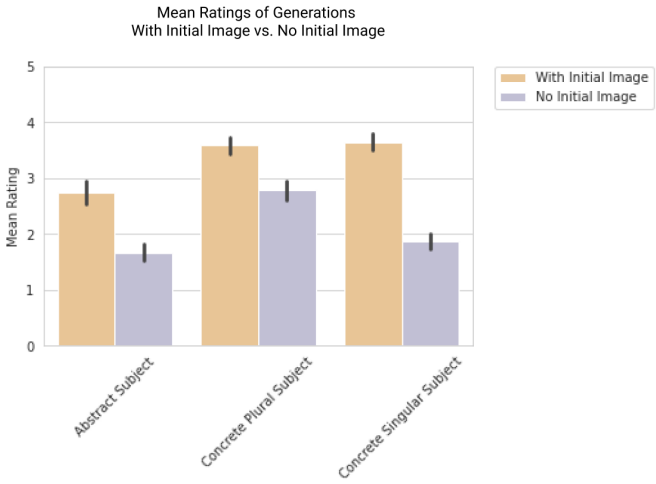

Figure 2: Average ratings of subject representation quality for each type of subject (abstract, concrete plural, concrete singular). Ratings for generations with initial images were significantly higher for all subject types, with the greatest improvement in the concrete singular subject type. Black bars show standard error.

We found that concrete plural subjects performed the best with or without a initial image. Although the difference between mean ratings of generations with and without initial images was statistically significant for concrete plural subjects, concrete plural subjects were already rated relatively highly (at 2.79 compared to 1.65 and 1.85 for abstract and concrete singular subjects respectively). We depict examples of the effect of initial images for concrete plural subjects in Figure 4. In the bottom group of Figure 4, we show generations of the subject road. Notice that all images generated without a initial image tended to show a similar composition: with a road converging towards a vanishing point. The rows of generations using initial images, however, captured different perspectives of the subject road. For example, one icon initial image produced a wavy road to represent the subject while another photo initial image produced a birds eye view of complicated road traffic interchanges. In the top group of Figure 4, we show another group of generations for the subject fire. While generations without initial images for fire were already rated highly, initial images nonetheless helped condition the way the subject was represented.

Table 1: Subject Representation Rubric for Annotators

Score	Description
5	Excellent representation of the subject, a very high number of features are present.
4	Good representation of the subject, high numbers of features are present.
3	Average representation of the subject, some features are present.
2	Bad representation of the subject, few features are present.
1	Extremely poor representation of the subject, subject not recognizable.

Table 2: Mean Rating by Subject Category

Abstract Subject		Concrete Plural Subject		Concrete Singular Subject	
2.19		3.18		2.75	
Initial	No Initial	Initial	No Initial	Initial	No Initial
2.73***	1.65***	3.58***	2.79***	3.64***	1.85***

*** indicates p-value <= 0.001

Abstract subjects received the lowest mean ratings in both cases: with and without initial images. Often, in low rated generations made without initial images, the model failed to find visual representations of these abstract concepts. Even when initial images were used to prompt the model, the model tended to misinterpret some elements in the initial images. For example, generations in the subject of "hate" are depicted in Figure 5 (bottom). Elements relevant to hate can hardly be identified in generations made without initial images. For generations of hate using an icon initial image of two people arguing, the model failed to recognize the two people and turned the lines into text. Similarly, when generations of hate used a photo initial image, the original photo was transformed into something in disarray and far apart from the original image. However, certain subjects performed better when the model was able to find a symbolic representation of the abstract subject. For example, generations of love were rated much higher, because the model was able to use the heart symbol to represent love. While the model constantly defaulting to creating heart shapes is not necessarily wrong, we found that initial images could help steer the generation towards other symbolic representations of love (i.e. a hug seen in the bottom two rows of Figure 5).

From these results, we synthesize the following design guidelines:

- **For concrete singular subjects, use initial images to improve the coherence of the subject in the generation.**
- **For concrete plural subjects, use initial images to draw out different perspectives and composition.**
- **For abstract subjects, try a variety of initial images to steer the model towards recognizable symbols.**

3.3.3 RQ 2 Quantitative Analysis: What types of initial images will lead to better subject representation?

While we concluded that initial images can improve subject representation in generations, we wanted to understand what qualities of initial images can benefit the user more. For example, is it more important to have larger, salient details or fine-grained details? Do generations have a difference in quality when the initial image is more stylized and symbolic or when the initial image is more photorealistic? It is important for designers to understand the level of detail and degree of freedom they have in the initial images, because there are countless images that could be utilized as prompts. To address these questions, we investigate whether icon initial images and photo initial images significantly differ in generation quality across abstract, concrete singular, and concrete plural subjects.

To answer *RQ2: What types of initial images, icon or photo, will lead to better subject representation across different types of subjects*, we use Mann-Whitney U tests to compare the mean ratings of generations with icon initial images and mean ratings of generations with photo initial images. To test H2.1, we used a Mann-Whitney U test to compare distribution of the mean ratings of generations with icon initial images and generations with photo initial images in the abstract subject category. We found that the mean ratings of generations with icon initial images is significantly different from the mean ratings of generations with photo initial images (p-value =0.03). The mean rating of generations with icon initial images is 2.96 and the mean rating of generations with photo initial images is 2.51. The result differs from our original hypothesis H2.1. Originally, we thought photos would help produce more realistic and detailed depictions of abstract concepts, but in the later qualitative analysis section, we will discuss how information from photo initial images were erased by the model and how information from simple icons can be preserved. To test H2.2, we again used a Mann-Whitney U test to compare the mean rating of generations with icon initial images and generations with photo initial images in the concrete plural subject category. We found no significant difference, with a mean rating of 3.62 for generations with icon initial images and a mean rating of 3.54 for generations with photo initial images. This result again differs from our original hypothesis H2.2 which was formed with the same reasonings as stated above. Lastly, we conducted a Mann-Whitney U test for H2.3, comparing the mean ratings of generations with icon initial images and photo initial images in the concrete singular subject category. The difference in the distribution of ratings was again insignificant, with a mean rating of 3.56 for generations with icon initial images and a mean rating of 3.73 for generations with photo initial images. This result aligns with our original hypothesis. Table 3 and Figure 6 show and visualize the mean ratings for each combination of subject category with icon and photo initial images.

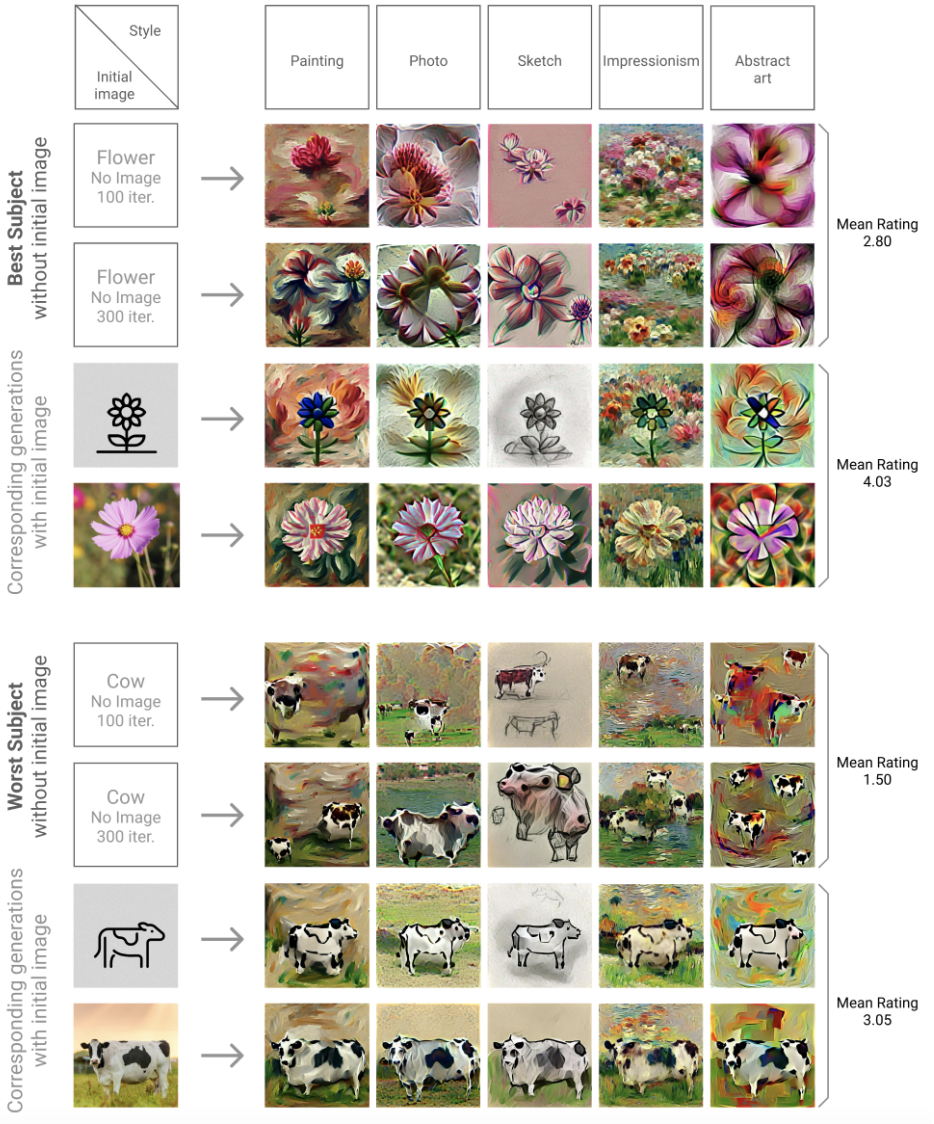

Figure 3: Best and worst examples for concrete singular subjects with and without initial images. The subject flower without initial images had the highest mean ratings at 2.80. The subject flower with initial images had a mean rating of 4.03 (top). The subject cow without initial images had the worst mean ratings at 1.50. The subject cow with initial images had a mean rating of 3.05 (bottom). Initial images consistently improved subject representation for concrete singular subjects, even for the best subject.

Table 3: Mean Rating by Subject Category

Abstract Subject (with initial) 2.73		Concrete Plural Subject (with initial) 3.58		Concrete Singular Subject (with initial) 3.64	
Icon	Photo	Icon	Photo	Icon	Photo
2.96*	2.51*	3.62	3.54	3.56	3.73

* indicates p-value <= 0.05

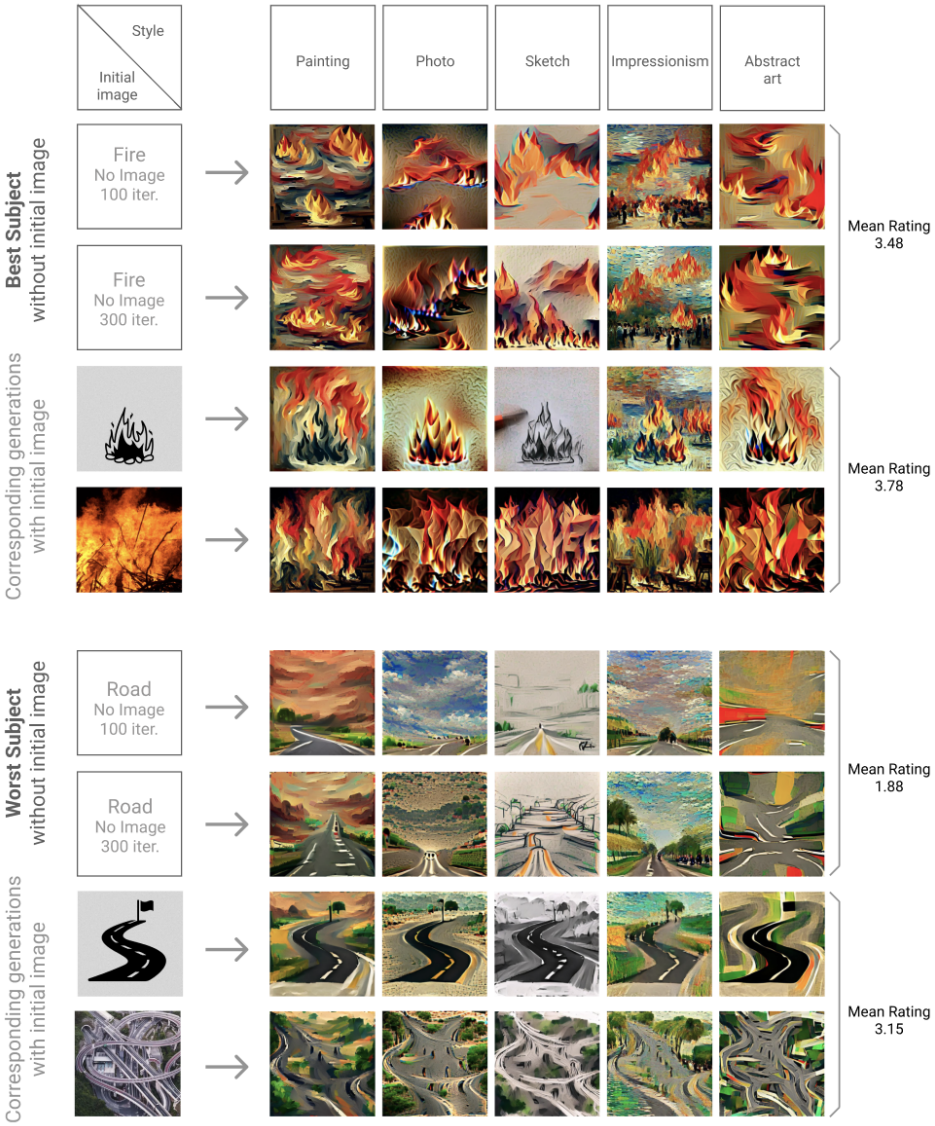

Figure 4: Best and worst examples for concrete plural subjects with and without initial images. The subject fire without initial images had the highest mean ratings at 3.48. The subject fire with initial images had a mean rating of 3.78 (top). The subject road without initial images had the worst mean ratings at 1.88. The subject road with initial images had a mean rating of 3.15 (bottom). Initial images consistently improved subject representation for concrete plural subjects, but the best concrete plural subjects performed equally well with or without initial images.

3.3.4 RQ 2 Qualitative Analysis: What types of initial images will lead to better subject representation?

For abstract subjects, ratings of generations using icon initial images were significantly higher than those of generations using photo initial images. As discussed in the previous section, abstract subjects were rated the lowest among all subject categories because the model failed to generate relevant symbols and produce recognizable elements from the initial images. We found that high rated abstract subject and icon combinations benefited from the minimal but highly salient details that make icons what they are. Even though these minimal details were simplified representations of abstract concepts, they were easier to preserve throughout the model's optimization than complex and realistic photos. The salience and easily interpretable nature of the details also made it so even if a part of an icon is distorted or erased by the model, the subject can still be well represented. For example, in Figure 7, the sad face from the icon was consistently preserved across the icon generations of sadness, whereas many details of the photo initial images relevant to sadness were erased in the photo generations, making it hard to recognize the subject.

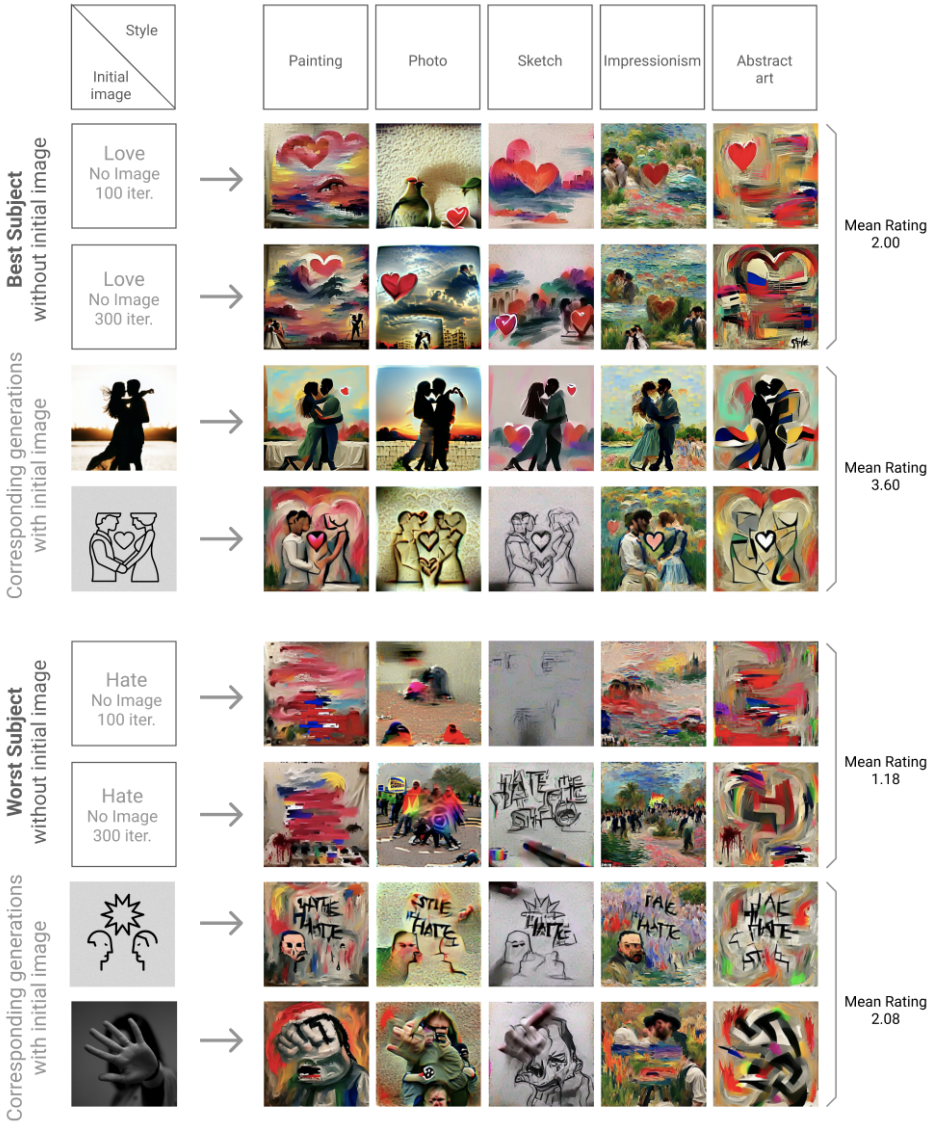

Figure 5: Best and worst examples for abstract subjects with and without initial images. The subject love without initial images had the highest mean ratings at 2.00. The subject love with initial images had a mean rating of 3.60 (top). The subject hate without initial images had the worst mean ratings at 1.18. The subject hate with initial images had a mean rating of 2.08 (bottom). Initial images consistently improved subject representation for concrete plural subjects, even for the best subject.

We also conducted a Mann-Whitney U test for each of the five art styles, painting, photo, sketch, impressionism and abstract art, to investigate whether the ratings of photo initial generations and icon initial generations differed across art styles. We found mean ratings of icon initial generations were significantly higher than photo initial generations for the abstract art style (p-value = 0.005). For the other art styles, with p-values greater than 0.1, the difference between icon and photo initial images was not significant. Table 4 and Figure 8 show and visualize the mean ratings for each combination of art styles with icon and photo initial images.

Since the difference in ratings was significant between photo and icon initial images in the abstract art style category, the best and worst examples are presented in Figure 9 below. All images with mean rating equal to 1 are generated with photo initial images and all images with mean rating equal to 5 are generated with icon initial images. The highly rated generations using icon initial images all have a clear subject sitting on top of a background in the style of abstract art. The generations with lowest ratings have a deconstructed quality that make the subject slightly indistinguishable from the background.

Table 4: Mean Rating by Art Style

Painting		Photo		Sketch		Impressionism		Abstract Art	
3.40		3.38		3.45		3.31		3.05	
Icon	Photo	Icon	Photo	Icon	Photo	Icon	Photo	Icon	Photo
3.40	3.40	3.29	3.46	3.42	3.49	3.38	3.25	3.40**	2.69**

** indicates p-value <= 0.01

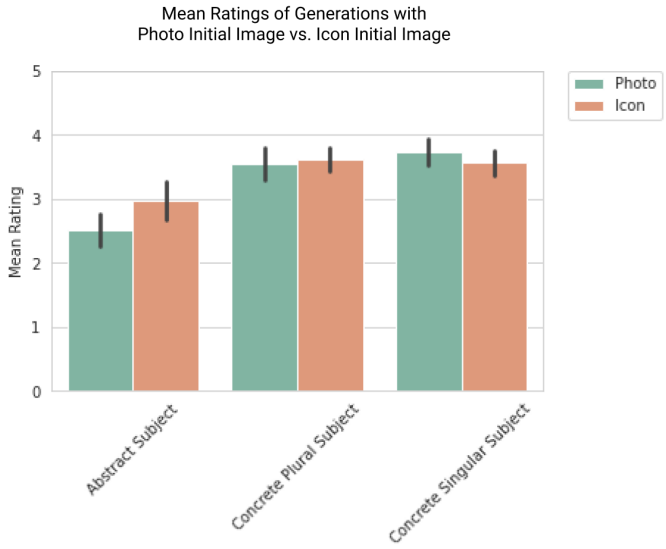

Figure 6: Average ratings of subject representation quality for each type of subject (abstract, concrete plural, concrete singular). For abstract subject types, ratings for generations with icon initial images were significantly higher than generations with photo initial images. For other types, there is no significant difference. Black bars show standard error.

We found that the aesthetics of the generations with initial images depend on the aesthetics of the initial images. As shown in Figure 10, many of the generations with icon initial images take on the stylization of the icon, which is flat and cartoonish; and generations with photo initial images show more realistic depiction of the subject. It is noteworthy that although the stylization of icons is preserved, the black and white nature of the icon is replaced with relevant colors according to the subject and the style in the text prompts.

From the previous result, we synthesis the following design guidelines:

- **For abstract subjects, using icon initial images produce more apparent subject representation.**
- **Using a variety of initial images is valid for concrete singular subjects and concrete plural subjects. There is no difference between the quality of generations with icon initial image and photo initial images.**

4 DISCUSSION

In this study, we looked at the interaction of subject and initial image type on the quality of generation. We found that initial images significantly improved subject representation across abstract, concrete singular, and concrete plural subjects. Furthermore, we found that icons and photos, different types of initial images expressing different extremes of detail and stylization, retrieved better or worse outcomes from the model depending upon the art style and subject type. The design guidelines are reiterated below, and we discuss the implications of these guidelines and findings in the following sections.

- For concrete singular subjects, use initial images to improve subject coherence.
- For concrete plural subjects, generating without an initial image can lead to high quality results, but use initial images to increase variety of perspectives and global composition.
- For abstract subjects, try generating without an initial image, as that may lead to decent outcomes. Then try a variety of initial images to help the model find recognizable symbols.
- For abstract subjects, use icon initial images to produce more apparent subject representation.
- For concrete singular and concrete plural subjects, try a variety of initial images of different levels of detail and stylization. We found no difference between the quality of generations made from icon initial images than those made from photo initial images.

4.1 Balancing Text and Image Information in the Prompt

Our results suggest that generations were rated higher when the text prompt and initial image complemented each other. For example, we found that concrete singular subjects were significantly improved when they were generated with photo initial images, because these subjects benefited from the natural structure of the photos. Generations of abstract art styles paired with icon initial images also did well, potentially because the minimal yet salient details of the icons left more room for the stylistic cues of abstract art to come through. When information from both the text and image prompt were aligned, both modes contributed to the final generation aesthetic.

However, sometimes the model was unable to incorporate information from both modes. Examples of this were seen in generations in the "Impressionism" and "painting" styles. The style dominated the aesthetic of these generations, while details from the initial images tended to drop out. For example, if we used a photo initial image, the 3D perspective and original colors of the photo would recede into the generation, which would default to a stereotypical

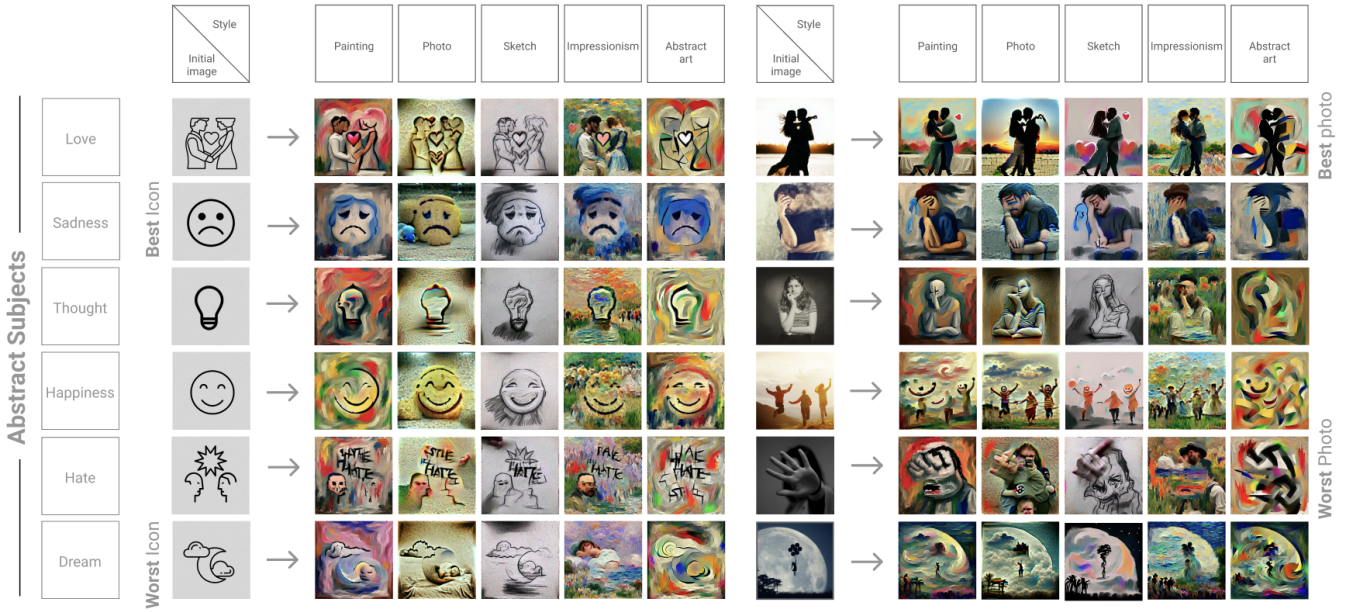

Figure 7: Generations for six abstract subjects using both types of initial images: icon and photo. For abstract subjects, icon initial images (left) received significantly higher ratings than photo initial images (right).

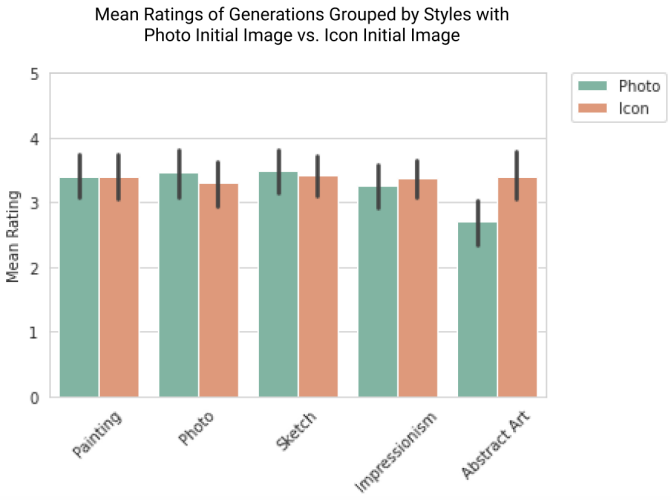

Figure 8: Mean Ratings for different art styles. Average ratings of subject representation quality for each art styles (painting, photo, sketch, impressionism, abstract art). For abstract art, ratings for generations with icon initial images were significantly higher than generations with photo initial images. For other art styles, there is no significant difference. Black bars show standard error.

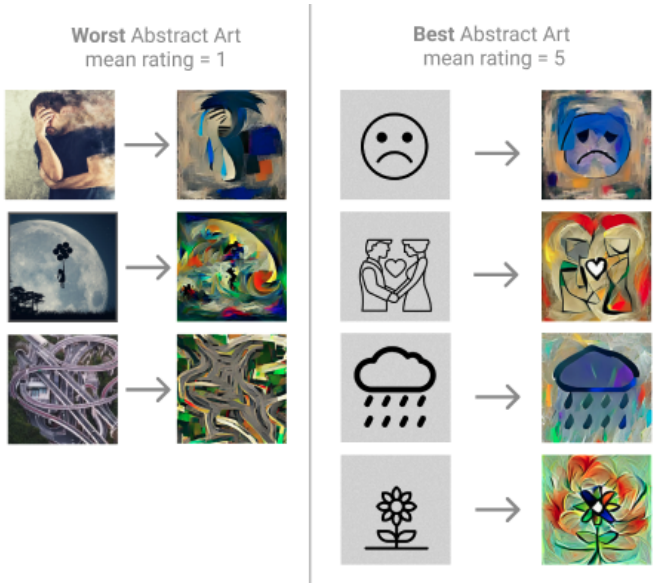

Figure 9: Best and worst generations with initial images in the abstract art style. All generations in the worst group have photo initial images (mean rating = 1). All generations in the best group have icon initial images (mean rating = 5).

Impressionist or painting style. This suggests that information from the text and image prompts can be conflicting and hard to resolve. Lastly, some last examples shown in Figure 12 illustrate scenarios where the model was unable to correctly incorporate the text and image sides of the prompt; the model misinterpreted initial image details as visualized text.

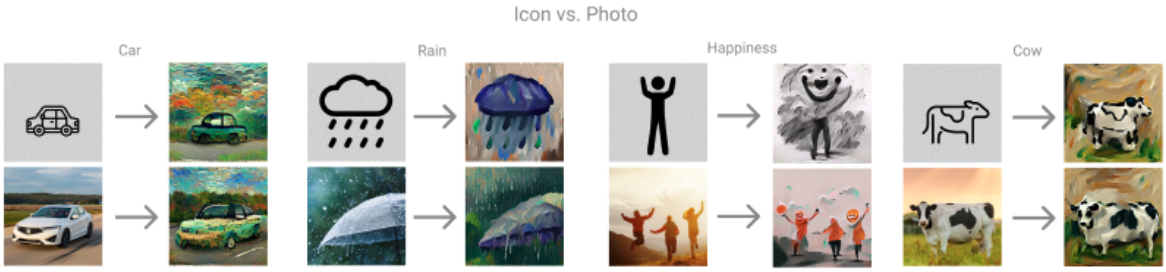

Figure 10: Examples of images generated from icon and photo initial images using the same text prompt. Icon and photo initial images lead to different aesthetics.

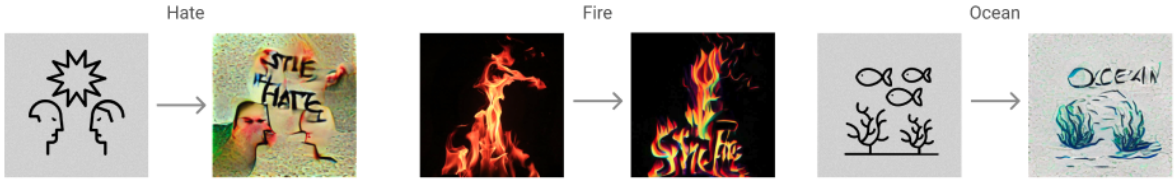

Figure 11: Examples of model misinterpreting initial image details to visualize texts.

4.2 Generating Abstract and Representational Art

In our related work, we described a body of work investigating the ways we cognitive process abstract and representational art. However, text-to-image generations often fail to fall cleanly into either category because they illustrate too many uncanny features. Perspectives and compositions often do not follow natural structures like that of representational art. Generations often also lack the meaning and emotional power of color that are embedded in artist-created abstract art. Our work provides design guidelines to improve generations along both of these traditions. However, there may be a limit to how much we can make these generations analogous to abstract or representational art. A painting of "a sunset in the style of Impressionism" prompted with a photo initial image of the Seine River is debatably not an Impressionist painting because it is an anachronistic artifact that lacks the spirit of the movement–something designed by an AI, not a human.

This suggests that while we need to understand how to generate visual media such that it respects what we are cognitively used to processing, we should also think about how the divide between abstract and representational art changes if we think about AI generated art as a new art form. For the first time, the human is not responsible for specifying every detail from corner to corner of the canvas. One analogue to help express this would be to compare the role architects fulfill in design compared to the role gardeners fulfill. Perhaps in AI generated art, the human shift from an architect-aligned role to a gardener-aligned role, responding to the plurality of options before them, pruning for preference as opposed to specifying full control.

4.3 Future Work & Limitations

While our design guidelines were intended to scaffold a wide variety of potential artwork possibilities, end users could benefit from still more fine-grained design guidelines. Many of the generations tended to have an uncanny quality to them, because they did not have consistent perspective, technique, composition, or color palette throughout. How could a designer control the perspective, details, or color palette of the final generation? For example, investigating how to maintain a wholly 2D or 3D perspective throughout a generation by using perspective cues in initial images could produce design guidelines that reduce the type of uncanniness specific to conflicting perspectives.

Another open question is how certain details within the initial images can be maintained during the optimization process. Often, certain details that are smaller within the generations can be lost and blended away as optimization goes on. For example, if we look at the photo initial images for "dream" in Figure 7, we can see that the small details such as the balloons of the photo initial image are not preserved as intended. It would be valuable for a user to understand what details are less useful to provide within an initial image. Additionally, it would be worth investigating how users can add details through multiple rounds of generation, thereby understanding how users can be supported through iteration.

We could also study a wider variety of initial image forms. In this study, we focused on icons, a simple and symbolic way to represent a subject and photo, a realistic and detailed way to represent a subject. In between, there are so many kinds of images, from vector illustrations to pixel art, which could provide more details than icons but less complexity than a photo. There are also many images made from simple shapes and colors that might be more in line with what users might manually draw or design for themselves and try as an initial image. It would be valuable to explore how the

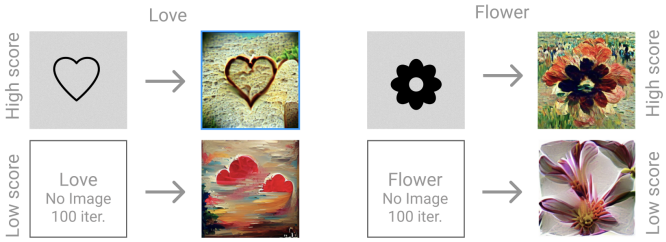

Figure 12: The top row shows generations that received very high ratings (mean ratings of 4.5), even though they may not be the most aesthetically pleasing. The bottom row shows generations received low ratings (mean ratings of 1.5), but they might be more aesthetically pleasing. This figure shows that improving subject representation does not mean making the image more aesthetically pleasing.

model responds to these types of images, so that they can utilize images of their own creation rather than being limited by the stock images online.

Lastly, future system work could study creativity support tools that integrate text-to-image generation. Systems could be made to study how users can efficiently explore an enormous amount of design solutions made from multiple types of initial images and text prompts. Generative AIs could present as alternative image search engines which retrieve thousands of images for users to take, composite, and remix.

We conclude by acknowledging a few limitations of our study. One was that we annotated these generations based upon how the subject was represented in the image. However, we did not look at the aesthetic value of these generations. In some generations, the subject came through very saliently, even though the generation lacked the aesthetic qualities that would make it usable for designers. Figure 12 shows examples of generations that received high ratings but might not be more aesthetically pleasing than the generations received low ratings.

A second limitation of this study is that all the icons we used in our study are 2D icons. There are 3D icons that keep the symbolic and simple aesthetics of a traditional 2D icon but they show more depths and perspectives. We did not encounter any of these icons when collecting data for our experiment, but we want to acknowledge the different effects that 2D and 3D icons may have on the generation outcomes.

5 CONCLUSION

In this paper, we conducted an annotation experiment to answer when initial images help improve the subject representation in AI generated images and what type of initial images we should use with different text prompts. The experiment results indicate that using an initial image significantly improves the quality of subjects representation in generations across all three subject categories (abstract, concrete plural and concrete singular) and that icon initial images are significantly better than photo initial images at presenting high quality subjects in the abstract category. In addition, we summarized in qualitative analysis ways different types of initial images can improve generation quality as well as produce different aesthetics in generations. We integrate our findings into design guidelines that can scaffold the text-to-image generative process for users and improve the control of art outcomes from multimodal AI.

ACKNOWLEDGMENTS

This paper is supported by NSF grant DGE - 1644869.

REFERENCES

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. 2018. Learning Representations and Generative Models for 3D Point Clouds. arXiv:1707.02392 [cs.CV]
[2] Adverb. 2021. Advadnoun. https://twitter.com/advadnoun
[3] Gunjan Aggarwal and Devi Parikh. 2020. Neuro-Symbolic Generative Art: A Preliminary Study. arXiv:2007.02171 [cs.AI]
[4] Refik Anadol. 2019. Latent History. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 1138. https://doi.org/10.1145/3343031.3355700
[5] M Augustin, Helmut Leder, Florian Hutzler, and Claus-Christian Carbon. 2008. Style follows content: On the microgenesis of art perception. *Acta psychologica* 128 (06 2008), 127–38. https://doi.org/10.1016/j.actpsy.2007.11.006
[6] Vered Aviv. 2014. What does the brain tell us about abstract art? *Frontiers in Human Neuroscience* 8 (2014). https://doi.org/10.3389/fnhum.2014.00085
[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096 [cs.LG]
[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] Kate Compton and Michael Mateas. 2015. Casual Creators. In *ICCC*.
[10] 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
[11] Gerald Cupchik, Oshin Vartanian, Adrian Crawley, and David Mikulis. 2009. Viewing artworks: Contributions of cognitive control and perceptual facilitation to aesthetic experience. *Brain and Cognition* 70 (06 2009), 84–91. https://doi.org/10.1016/j.bandc.2009.01.003
[12] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. 2021. DALLE Mini. https://doi.org/10.5281/zenodo.1234
[13] Kevin Frans, L. B. Soros, and Olaf Witkowski. 2021. CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders. https://doi.org/10.48550/ARXIV.2106.14843
[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. A Neural Algorithm of Artistic Style. arXiv:1508.06576 [cs.CV]
[15] Songwei Ge and Devi Parikh. 2021. Visual Conceptual Blending with Large-scale Language and Vision Models. arXiv:2106.14127 [cs.CL]
[16] 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]
[17] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]
[18] Eric Kandel. 2016. *Reductionism in Art and Brain Science: Bridging the Two Cultures* (1st ed.). Columbia University Press, USA.
[19] Pegah Karimi, Kazjon Grace, Mary Lou Maher, and Nicholas Davis. 2018. Evaluating Creativity in Computational Co-Creative Systems. https://doi.org/10.48550/ARXIV.1807.09886
[20] 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]
[21] Hideaki Kawabata and Semir Zeki. 2004. Neural correlates of beauty. *Journal of neurophysiology* 91 4 (2004), 1699–705.
[22] Vivian Liu and Lydia B. Chilton. 2021. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. arXiv:2109.06977 [cs.HC]
[23] Maria Teresa Llano, Mark d'Inverno, Matthew John Yee-King, Jon Mccormack, Alon Ilsar, Alison Pease, and Simon Colton. 2020. Explainable Computational Creativity. In *ICCC*.
[24] Paul Locher, Elizabeth Krupinski, Claudia Mello-Thoms, and Calvin Nodine. 2007. Visual interest in pictorial art during an aesthetic experience. *Spatial vision* 21 (02 2007), 55–77. https://doi.org/10.1163/156856807782753868
[25] Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George Fitzmaurice. 2018. Dream Lens: Exploration and Visualization of Large-Scale Generative Design Datasets. In *Proceedings of the 2018 CHI Conference on*

Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173943

[26] Alexander Mordvintsev, Michael Tyka, and Christopher Olah. 2015. google/deepdream. https://github.com/google/deepdream

[27] 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

[28] Frieder Nake. 2007. Computer Art: Creativity and Computability. In *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition* (Washington, DC, USA) *(C&C '07)*. Association for Computing Machinery, New York, NY, USA, 305–306. https://doi.org/10.1145/1254960.1255041

[29] nerdyroden. 2022. nerdyrodent/VQGAN-CLIP. https://github.com/nerdyrodent/VQGAN-CLIP

[30] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arXiv:2103.17249 [cs.CV]

[31] Elena Petrovskaya, Christoph Sebastian Deterding, and Simon Colton. 2020. Casual Creators in the Wild : A Typology of Commercial Generative Creativity Support Tools. https://eprints.whiterose.ac.uk/160760/

[32] 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]

[33] Janet Rafner, Arthur Hjorth, Sebastian Risi, Lotte Philipsen, Charles Dumas, Michael Mose Biskjær, Lior Noy, Kristian Tylén, Carsten Bergenholtz, Jesse Lynch, Blanka Zana, and Jacob Sherson. 2020. *Crea.Blender: A Neural Network-Based Image Generation Game to Assess Creativity.* Association for Computing Machinery, New York, NY, USA, 340–344. https://doi.org/10.1145/3383668.3419907

[34] Evan Shimizu, Matthew Fisher, Sylvain Paris, James McCann, and Kayvon Fatahalian. 2020. Design Adjectives: A Framework for Interactive Model-Guided Exploration of Parameterized Design Spaces. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 261–278. https://doi.org/10.1145/3379337.3415866

[35] Richard Taylor, Branka Spehar, Caroline Hagerhall, and Paul Van Donkelaar. 2011. Perceptual and Physiological Responses to Jackson Pollock's Fractals. *Frontiers in Human Neuroscience* 5 (2011). https://doi.org/10.3389/fnhum.2011.00060

[36] Oshin Vartanian and Vinod Goel. 2004. Neuroanatomical correlates of aesthetic preference of paintings. *Neuroreport* 15 (05 2004), 893–7. https://doi.org/10.1097/00001756-200404090-00032