ArtWhisperer: A Dataset for Characterizing Human-AI Interactions in Artistic Creations

Kailas Vodrahalli [1] **James Zou** [1]

Abstract

In this work, we investigate how people use text-to-image models to generate desired target images. To study this interaction, we created ArtWhisperer, an online game where users are given a target image and are tasked with iteratively finding a prompt that creates a similar-looking image as the target. Through this game, we recorded over 50,000 human-AI interactions; each interaction corresponds to one text prompt created by a user and the corresponding generated image. The majority of these are repeated interactions where a user iterates to find the best prompt for their target image, making this a unique sequential dataset for studying human-AI collaborations. In an initial analysis of this dataset, we identify several characteristics of prompt interactions and user strategies. People submit diverse prompts and are able to discover a variety of text descriptions that generate similar images. Interestingly, prompt diversity does not decrease as users find better prompts. We further propose a new metric to quantify AI model *steerability* using our dataset. We define steerability as the expected number of interactions required to adequately complete a task. We estimate this value by fitting a Markov chain for each target task and calculating the expected time to reach an adequate score. We quantify and compare AI steerability across different types of target images and two different models, finding that images of cities and nature are more steerable than artistic and fantasy images. We also evaluate popular vision-language models to assess their image understanding and ability to incorporate feedback. These findings provide insights into human-AI interaction behavior, present a concrete method of assessing AI steerability, and demonstrate the general utility of the ArtWhisperer dataset.

[1]Stanford University. Correspondence to: Kailas Vodrahalli <kailasv@stanford.edu>, James Zou <jamesz@stanford.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Direct human interaction with AI models has become widespread following a number of technical innovations improving the quality of text-to-text (Brown et al., 2020; Ouyang et al., 2022; Anil et al., 2023) and text-to-image models (Rombach et al., 2022a; Ramesh et al., 2022), enabling the public release of high-quality AI-based services like ChatGPT (chatGPT), Bard (Bard), and Midjourney (Midjourney). These models have seen rapid interest and adoption largely due to the ability of the general public to interact with and steer the AI in diverse contexts including engineering, creative writing, art, education, medicine, and law (Dakhel et al., 2023; Nguyen & Nadi, 2022; Ippolito et al., 2022; Cetinic & She, 2022; Qadir, 2023; Cascella et al., 2023; Sloan, 2023).

A key challenge in developing these models is aligning their output to human inputs. This is made challenging by the broad domain of use cases as well as the diverse prompting styles of different users. Many approaches can be categorized as "prompt engineering," where specific strategies for prompting are used to steer a model (Oppenlaender, 2022; Liu & Chilton, 2022; Zhou et al., 2022b; Wei et al., 2022; White et al., 2023). Great success has also been found by fine-tuning models using relatively small datasets to follow human instructions (Ouyang et al., 2022), respond in a specific style (Hu et al., 2021), or behave differently to specified prompts (Zhou et al., 2022a; Gal et al., 2022).

In this work, we take interest in the fact that human interaction with these models is often an iterative process. We develop a dataset to study this interaction. The dataset is collected through an interactive game we created where players try to find an optimal prompt for a given task (see Figure 1). In particular, we focus on text-to-image models and ask the player to generate a similar image (*AI Image*) to a given *target image*. The player is allowed to iterate on their prompt, using the previously generated image(s) as feedback to help them adjust their prompt. A score is also provided as feedback to help the user calibrate how "close" they are to a similar image.

Using this setup, we collected data on 51,026 interactions from 2,250 players across 191 unique target images. The target images were selected from a diverse set of AI-generated

and natural images. We also collected a separate dataset of 4,572 interactions, 140 users, and 51 unique target images in a more controlled setting to assess the robustness of our findings.

Based on this data, we find several interesting patterns in how people interact with AI models. Players discover a diverse set of prompts that all result in images similar to the target. To discover these prompts, players typically make small, iterative updates to their prompts. Each update improves their image with a moderate success rate ($40-60\%$ for most target images). Based on these findings, we define and evaluate a metric for model steerability using the stopping time of an empirical Markov model. We use this metric to assess steerability across image categories and across two AI models.

Our contributions We release a public dataset on human interactions with an AI model. To our knowledge, this is the first such dataset showing repeated interactions of people with a text-to-image model to accomplish specified tasks. We also provide an initial analysis of this data and propose a simple-to-calculate metric for assessing model steerability. Additionally, we use our dataset and this steerability metric to evaluate the ability of vision-language models to utilize feedback. Our dataset and associated code is made available at https://github.com/kailas-v/ArtWhisperer.

Related Works Human-AI interaction datasets for text-to-text and text-to-image models typically focus on single interactions and generally do not provide users with a specific task. Public text-to-image interaction datasets typically contain the generated AI images and/or prompts (Santana, 2022; Wang et al., 2022) and optionally some form of human preference rating (Pressman et al., 2022; Wu et al., 2023; Xu et al., 2023; Kirstain et al., 2023). These datasets generally rely on scraping online repositories like Lexica (Lexica) or Discord servers focused on AI art. Though some of these datasets include metadata that may allow for reconstruction of prompt iteration, there is no guarantee the user has the same desired output in mind over the iteration. Public text-to-text interaction datasets are much more limited as the best performing models are generally accessible only through APIs with no public user interaction datasets. While some researchers have investigated how human-AI interaction for text-to-text can be improved through various tools (Wu et al., 2022a;b), the amount of data collected is limited and not publicly available. There are also repositories containing prompt strategies for various tasks (Bach et al., 2022) but no human interaction component.

We seek to rectify two of the shortcomings of the existing datasets–namely, that they do not contain extended interactions as the user attempts to steer the AI, and they do not have a predefined goal. In our work, we maintain a controlled environment for human users where we allow extended interactions towards a fixed goal. As shown by our initial analysis, our dataset may enable deeper understanding of user prompting strategies, assessing model steerability, and evaluating vision-language AI models.

2. Interaction Game

In our game, players are shown a target image (see Figures 2 and 4). Players are also given a limited interface to a text-to-image model, Stable Diffusion (SD) v2.1 model (Rombach et al., 2022b). In particular, players can enter a "positive prompt" (describes the desired content) and a "negative prompt" (describes what should be omitted) to steer the AI model. All models hyperparameters including the seed are fixed. Upon inputting a prompt, the player is shown the image generated by the AI model, along with a similarity score between their generated image and the target image. The interface is shown in Figure 1.

2.1. How Target Images are Selected

We randomly sample target images from two sources. The first is a collection of Wikipedia pages, and the second is a dataset of prompts AI artists have used with SD (Santana, 2022). In addition to sampling target images, we need to ensure the task is feasible to users. We do not allow users to adjust the seed or other parameters of the model, so we need to ensure the selected model parameters can generate reasonably similar images to the target image. We find that selecting an appropriate random seed is sufficient, and fix all other model parameters (see Appendix A.3 for examples of generated images and Appendix A.5 for details and discussion).

Wikipedia Images A collection of 35 Wikipedia pages on various topics including art, nature, cities, and various people. A full list of pages sampled from is provided in Appendix 2. From these pages, we scraped 670 figures licensed under the Creative Commons license. These figures were then filtered by which had captions, as well as which images were JPG or PNG images (i.e., not animated, and not PDF files), resulting in 557 images.

For each of the 557 images, we first resize and crop the image to size 512×512. The Wikipedia caption is used as the ground truth "prompt". Let the image-caption pair be denoted as (t_i, p_i^*). We sample the model on 50 random seeds, with p_i^* as the prompt input. This generates a set of 50 images: $S_i = \{(x_i, s_i) : i = 1, \ldots, 50\}$ for generated image x_i and seed s_i. Let $C(x)$ denote the CLIP image embedding (Radford et al., 2021) of an image x. Then we select the seed as s_{i^*}, where

$$i^* := \min_{i=1,\ldots,50} \left|\left| \frac{C(x_i)}{||C(x_i)||_2} - \frac{C(t_i)}{||C(t_i)||_2} \right|\right|_2 \tag{1}$$

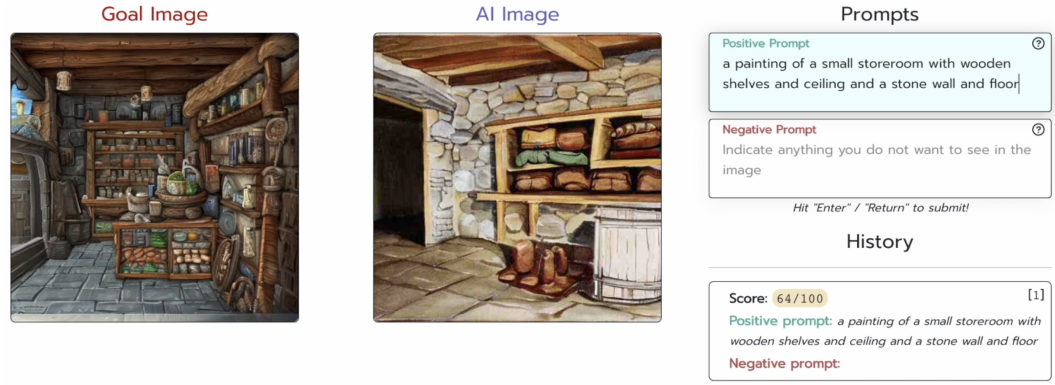

Figure 1. Interface of the ArtWhisperer game. Prompts entered on right. Target (goal) image and player-generated image on left. Previous prompts and scores are displayed in the lower right.

Here, s_{i^*} is selected to minimize the distance to the target image given the target prompt.

AI-Generated Images A collection of 2,000 AI-art prompts are randomly sampled from the Stable Diffusion Prompts dataset (Santana, 2022). For each prompt, p_i^*, we generate two sets of images. As before, we use 50 unique random seeds to select the seed, s_{i^*} and an additional 10 random seeds to use for selecting the generated target image (so in total, we use 60 unique random seeds): the first set, $S_{i,1} = \{(x_{i,1}, s_i) : i = 1, \ldots, 10\}$ and $S_{i,2} = \{(x_{i,2}, s_i) : i = 1, \ldots, 50\}$. We select the target image, $t_{i_1^*}$, from $S_{i,1}$:

$$i_1^* := \min_{i=1,\ldots,10} \text{median}\left(\left\{\left|\left|\frac{C(x_{i,1})}{||C(x_{i,1})||_2} - \frac{C(x_{j,2})}{||C(x_{j,2})||}\right|\right|_2 : j = 1, \ldots, 50\}\right) \quad (2)$$

We select the random seed, $s_{i_2^*}$, using $t_{i_1^*}$ and $S_{i,2}$, with

$$i_2^* := \min_{i=1,\ldots,50} \left|\left|\frac{C(x_{i,2})}{||C(x_{i,2})||_2} - \frac{C(t_{i_1^*})}{||C(t_{i_1^*})||_2}\right|\right|_2 \quad (3)$$

Here, $t_{i_1^*}$ is chosen to be more representative of the types of images we may expect given the fixed prompt, p_i^*. This is because $t_{i_1^*}$ is selected to be close to the center of the sampled images, $S_{i,2}$. The intuition for selecting $s_{i_2^*}$ is the same as selecting s_{i^*} for the Wikipedia images.

2.2. Scoring Function

To provide feedback to players, we created a scoring function to assess the similarity of a player's generated image and the target image. We define the scoring function as

$$score(x,t) = \max(0, \min(100, \alpha \cdot \left|\left|\frac{C(x)}{||C(x)||_2} - \frac{C(t)}{||C(t)||_2}\right|\right|_2 + \beta)) \quad (4)$$

for generated image x, target image t, and constants α, β. Note the range of $score(x,t)$ is integers in the interval $[0, 100]$. Details on how α, β are selected parameters are provided in Appendix A.4.

While this scoring function is often reasonable, it does not always align with the opinions of a human user. To assess how well $score(x,t)$ follows a user's preferences, we acquire ratings from a subset of users (see *ArtWhisperer-Validation* in Section 2.3). We find score(x,t) has a Pearson correlation coefficient of 0.579 indicating reasonable agreement. Further assessment is performed in Section 4.3 and discussed at length in Appendix A.14.

2.3. Dataset Overview

We collected two datasets: *ArtWhisperer* and *ArtWhisperer-Validation*. We use *ArtWhisperer* for most analysis and results; for some of the results in Sections 2.2, 4.2,and 4.3, we also use *ArtWhisperer-Validation* (when referenced). Data was collected from March-May 2023. IRB approval was obtained.

***ArtWhisperer*:** A public version of our game was released online, with three new target images released daily. We collected data from consenting users playing the game. Users were not paid. Users were anonymous and we only collected data related to the prompts submitted to ensure privacy of users. While we expect some users played the game across multiple days, we did not track them. A summary of the *ArtWhisperer* dataset is provided in Table 1. In total, we have 2,250 (potentially non-unique) players corresponding to 51,026 interactions across 191 target images. Players

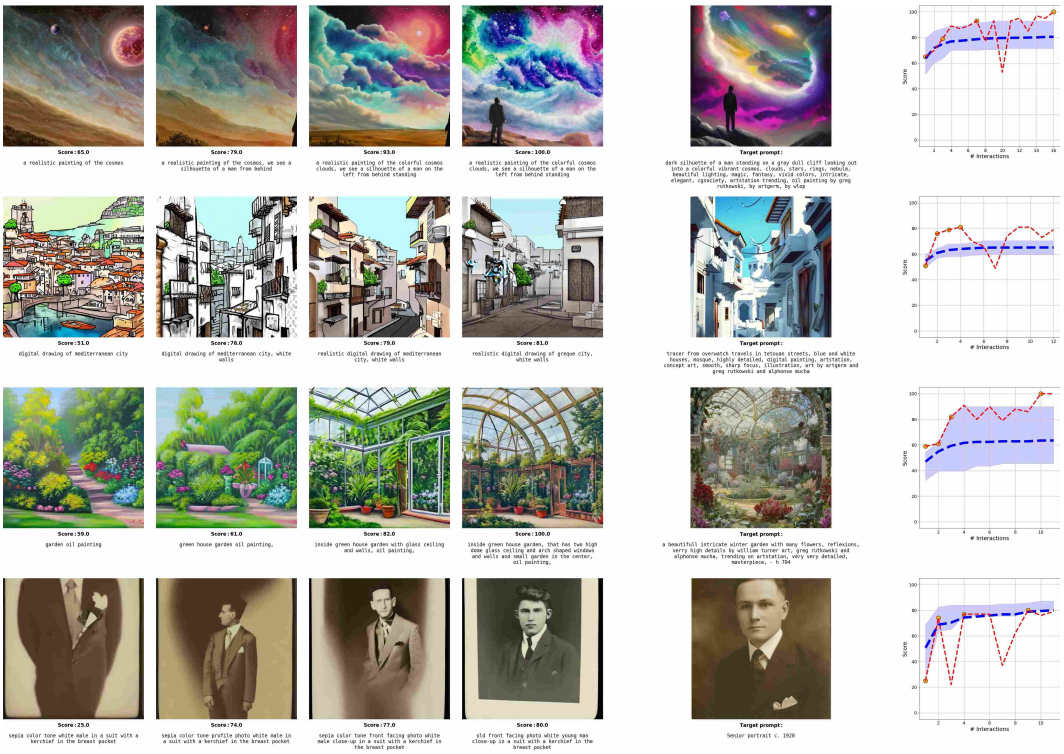

Figure 2. Example user trajectories. In each row, we show (1) a given user's prompts, (2) the target image (rightmost image), and (3) a plot of this target image's average score trajectory across users (blue), this user's full score trajectory (red), and the displayed images (orange).

interacted with the model SD v2.1. In Figure 3, we plot the number of queries submitted by players across different target images.

***ArtWhisperer-Validation*:** The game (with a near identical interface) was also released as a controlled user study to paid crowd workers on Prolific (Prolific Academic). The crowd workers were compensated at a rate of $12.00 per hour for roughly 20 minutes of their time. Workers played the game across 5 randomly selected target images from a pre-selected subset of 51 target images chosen to have diverse content. Workers were also asked to rate each of their images on a scale of 1-10 (i.e., self-scoring their generated images). In total, we collected data on 4,572 interactions, corresponding to 140 users and 51 unique target images across two different diffusion models, SD v2.1 and SD v1.5. Additional details and demographic information are provided in Appendix A.7.

3. Prompt Diversity

We quantify prompt diversity by looking at the distribution of prompts in the text embedding space. In particular, we use the CLIP text embedding (Radford et al., 2021), though we do find the choice of embedding is not particularly important for our results (see Appendix A.8).

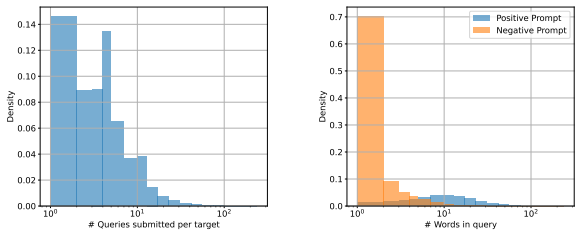

Figure 3. Left: Distribution of # of user queries per target image (average queries per image is 9.18). Right: Distribution of the # of words submitted in a query (average words submitted is 20.02 and 2.32 for positive and negative prompts respectively).

3.1. Diverse prompts used for high scores

First, we find that users do not converge in their prompt design but rather achieve similarly high scores through a diverse set of prompts. This result confirms the need for models to align to a wide range of user writing styles and descriptive techniques, and suggests a tradeoff between alignment to multiple users and specificity of a prompt (e.g., to support more users, individual prompts will necessarily have less descriptive ability). Examples of such diverse prompts are shown in Figure 4.

We quantify this finding in Figure 5, where we plot two

Table 1. ArtWhisperer Dataset Overview. Each row contains summary data for a different subset of the dataset. Subsets may overlap. Similar information for *ArtWhisperer-Validation* is in Appendix A.7.

# Players	# Target Images	# Inter-actions	Average # Prompts	Average Score	Median Duration	Category
2250	**191**	**51026**	**9.29**	**58.93**	**18 s**	**Total**
377	25	3884	8.65	56.70	19 s	Contains famous person?
353	32	3785	8.26	61.64	21 s	Contains famous landmark?
2005	140	40290	9.24	59.83	18 s	Contains man-made content?
1177	58	18255	10.93	57.21	17 s	Contains people?
344	77	6972	8.81	62.01	20 s	Is real image?
2140	103	43524	9.42	58.37	17 s	Is AI image?
1483	82	24913	9.14	59.45	17 s	Is art?
623	29	7297	9.14	53.77	18 s	Contains nature?
160	14	1355	7.28	65.74	19 s	Contains city?
1239	39	15872	9.91	56.74	16 s	Is fantasy?
618	19	8359	10.51	57.88	17 s	Is sci-fi or space?

metrics defined as follows. Let z_0, z_n be normalized embeddings of the initial and best prompt/image found by a user. Let z^* be the normalized embedding of the target prompt/image. We define the difference in embedding distance to ground truth as $||z_n - z^*||_2 - ||z_0 - z^*||_2$. In blue, we use the CLIP text embeddings of the prompts; in orange, we use the CLIP image embeddings of the generated images. We note two findings here: (1) the metric applied to the image embeddings is guaranteed to be non-positive as the embedding distance is monotonically decreasing with the score, and (2) the metric applied to the text embeddings is apparently symmetric around 0, indicating that unlike the image embedding, distance in the text embedding space does not monotonically decrease with score. Together, these findings illustrate that users tend to discover diverse prompts and *do not converge in their prompt design*.

Additionally, we find the distribution of prompts does not converge. In the left of Figure 6, we plot the distribution of distances between the first prompt (in blue) and the last prompt (in orange) to the average prompt for the corresponding target image. Despite the average score improving from 51.9 to 70.3 (out of 100) indicating a significant improvement in score, prompt diversity does not significantly diminish. That is, users do not converge to similar prompts to achieve high scores. Similar analysis of the image embedding space suggests image diversity *decreases* (Figure 5).

3.2. People submit similar prompts throughout their interaction

Second, we find that people do not explore a wide range of prompt designs even when their initial prompts are not performing well. This suggests the lack of convergence in prompt design is inherent to the users' preference to describe an image. It also suggests that user initialization (i.e., the first submitted prompt) is critical, and that online personalization may be possible (to adapt to the user's writing style) given the more stable nature of an individual's prompt design for a given target image.

In the center of Figure 6, we plot the distribution of the standard deviation of prompts for users (blue) and for permuted users (orange). Permuted users are generated by sampling from all prompts for a given target image uniformly, using the same distribution of number of prompts as for real users. The gap between the two distributions shows that individuals do not randomly sample prompts each interaction, but base new prompts off of previously submitted prompts (p-value $< 10^{-10}$, t-test for independent variables). An analysis of how scores change between adjacent prompts shows that this strategy has a moderate success rate and improves the score $40 - 60\%$ of the time, with an average rate of 48.6% (note that score changes < 1 are counted as unchanged; this occurs 10.2% of the time).

3.3. People have similar prompt styles across images

Finally, we find that people have a measurable writing style or prompt design that appears across images; however, this prompt design is predominantly informed by the target image rather than any consistent style. This implies that online personalization *across different target images* may require a large number of images to be effective (across a small set of individual target images, the prompt design and writing style may still change significantly). This is in contrast to our earlier finding that suggests personalization for a given image may be possible given the relative stability of a user's prompt design for a given target image.

We quantify user style by computing the difference (in the

CLIP text embedding space) between the average prompt of a given user and the average prompt across all users for a given target image. To quantify style variation for a user, we then compute the standard deviation of the user style across the target images the user generated. In the right of Figure 6, we plot the distribution of user style variation for real users (blue) and permuted users (orange). Permuted users are generated by randomly sampling user styles. This allows us to test whether users have a consistent prompting style. We find users do indeed have specific styles of prompting (p-value $< 10^{-10}$, t-test for independent variables). However, the difference is seemingly not large, suggesting that while user style may a component to prompting, other factors related to the target image may be more important.

4. Model Steerability

Model steerability refers to the ability of a user to steer a model towards a desired outcome. Measuring steerability has great utility, as it enables quantitative tracking of the human usability of generative models. However, there is no current consensus on how to measure AI steerability. A common approach is to simply measure performance of a model on standardized dataset evaluations (Jahanian et al., 2019; OpenAI, 2023). While this can enable comparisons between tasks and models, this approach does not allow for the feedback loop present when humans interact with a model. Steerability can also be measured qualitatively based on user assessment of their experience interacting with the AI (Chung et al., 2022).

Here, we propose a simple yet informative measure to assess model steerability. This metric is general across model types and data modalities, and we are able to use it here to compare across image categories and models. We then analyze this measure across different subgroups of images and across two different Stable Diffusion models: SDv2.1 and the older SDv1.5 (Rombach et al., 2022a).

4.1. Measuring steerability

As discussed in Section 3.2, users typically engage with the model through clusters of similar prompts. They typically start with an initial base prompt and proceed to make multiple incremental modifications to it. We use this observation as a basis for creating a steerability metric. We define a Markov chain between scores. Each node is a score with edges connecting to the subsequent score. To make this tractable for empirical analysis, we bin scores into five groups: $[0, 20], [21, 40], [41, 60], [61, 80], [81, 100]$. We use the expected time taken to reach the last score bin, $[81, 100]$, as our steerability score (i.e., the stopping time to reach an adequate score).

For each target image, we calculate the empirical transition probability matrix between binned scores using all the players' data for that image. We then calculate the *steerability score* for the given target image by running a Monte Carlo simulation to estimate stopping time, as defined above. To assess steerability across a group of images, we average steerability score across all images in the group.

4.2. Analysis

In Figure 7, we plot the steerability score across image groups. Error bars show the standard error. For examples of steerability scores for individual images, see Appendix A.12. We find that images containing famous people or landmarks, real images (not AI generated), contain cities, or contain nature are the most steerable. AI-generated images, fantasy images, and images of human art are the least steerable. There are a few possible explanations. The model we are assessing here, SDv2.1, as well as its text encoder OpenCLIP, are trained on subsets of LAION5B (Schuhmann et al., 2022). The contents of LAION5B are predominantly real world images, indicating why these images may be more steerable (i.e., text describing these types of images may have a better encoding). Moreover, the prompts for AI-generated images and fantasy images generally include specific internet artists and/or art styles which may not be known to most users making achieving the desired target image more difficult. Another potential reason is the distribution of images chosen for each category. Clearly, there are "easier" and "more difficult" images in each category; part of the reason for smaller stopping time may be the sample of images chosen rather than the actual image category.

Using the *ArtWhisperer-Validation* data, we also compare steerability across two models: SDv2.1 and SDv1.5. Across most image categories, we observe a similar steerability. Images of nature, sci-fi or space, and real images have the largest differences in steerability between the two models; SDv2.1 is more steerable in all three cases. This suggests that SDv2.1 may be more steerable for natural images as well as sci-fi images, and is similarly steerable for other kinds of images including AI-generated artwork. One explanation may be that most of our users were not aware of certain prompting strategies that help models generate more aesthetic images or certain art styles; it is possible that for experienced users, AI art images may be more steerable, and differences between models may be magnified if, for example, a user is experienced working with one particular model. More discussion is provided in Appendix A.11.

4.3. Justification for automated score

One limitation of our steerability metric comes from the method of scoring user-submitted prompts. Ideally, we would assess steerability with a user's personal preferences. As mentioned in Section 2.2, the scores and human ratings

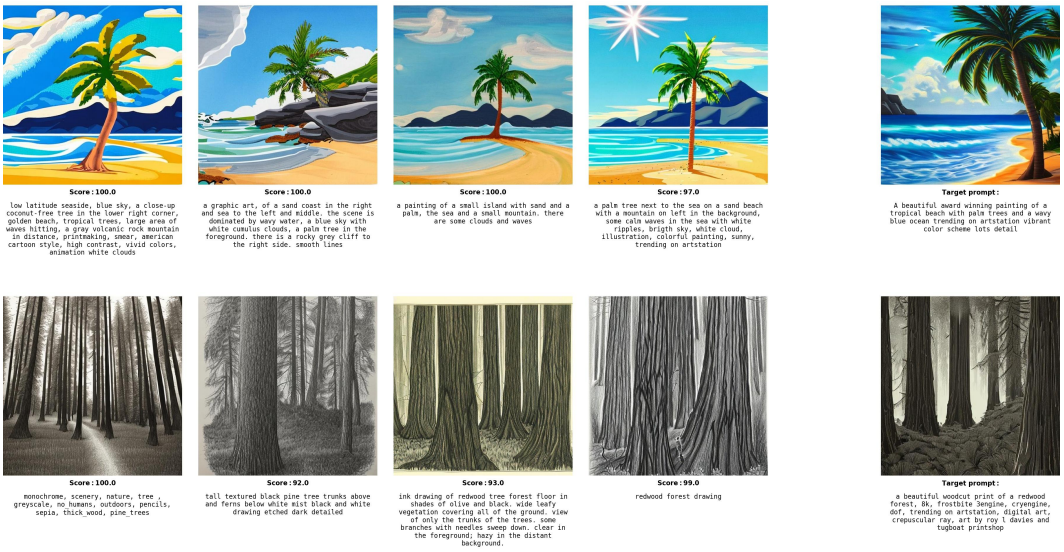

Figure 4. Diverse, high-scoring prompt submissions from different users. Target image in rightmost column.

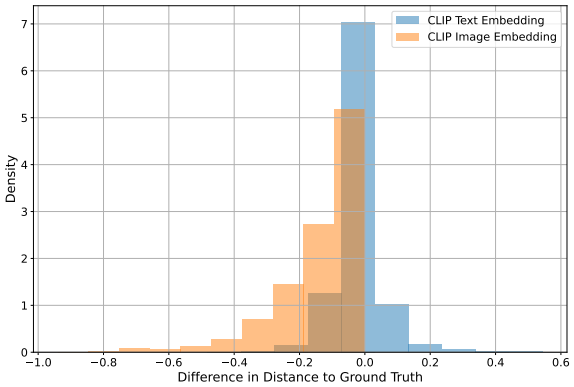

Figure 5. Difference of distance from the first prompt to ground truth and distance from the last (best) prompt to ground truth for CLIP text (blue) and CLIP image embeddings (orange).

have a positive correlation. Here, we use the human ratings from *ArtWhisperer-Validation* instead of our score function to assess steerability. We compute the steerability score across both models and across image groups. Generally, the steerability scores change little. In all but two cases (SDv2.1 on sci-fi and space images; SDv1.5 on nature images), the human rating-based steerability score remains within a 95% confidence interval of the score-based steerability score. While our score function may not perfectly capture human preferences, the steerability score we generate appears to be robust to these issues. Further discussion is included in Appendix A.14.

5. Vision-Language Model Evaluation

We also use the ArtWhisperer dataset to evaluate vision-language models (VLMs) for their ability to incorporate feedback. In particular, we have two VLMs, GPT-4 and Gemini (OpenAI, 2023; Team et al., 2023), play the ArtWhisperer game across all target images. In this context, the models are interacting with a "tool" – an SD model. However, we can also consider this SD model as a proxy for a user whose preferences the model must adapt to over time.

A system prompt is crafted to inform the model about the game. A starting prompt is used to query the model for an initial prompt. The generated prompt is evaluated using the ArtWhisperer game pipeline – an image is generated and then scored. In the "Feedback" mode, this generated image and the ArtWhisperer score is then fed back to the VLM with a request for an updated prompt. This process is repeated until the model attains a perfect score or 20 attempts have been made. In "No feedback" mode, the VLM is not given any feedback and is just queried repeatedly without any conversation history given. More details on the prompts used are included in Appendix A.16.

In Figure 8, we compare the average score (across target images) trajectory of the VLM models over time. Here, the score is the ArtWhisperer game score, where larger values indicate a better prompt. For each of GPT-4 and Gemini, we have two trajectories. "GPT-4" and "Gemini" represent the prompting methodology described above and further detailed in Appendix A.16. The "No feedback" plots represent the average score across queries as well, as each query is independent; this is a baseline measure of the VLM's ability to prompt Stable Diffusion given an image with no

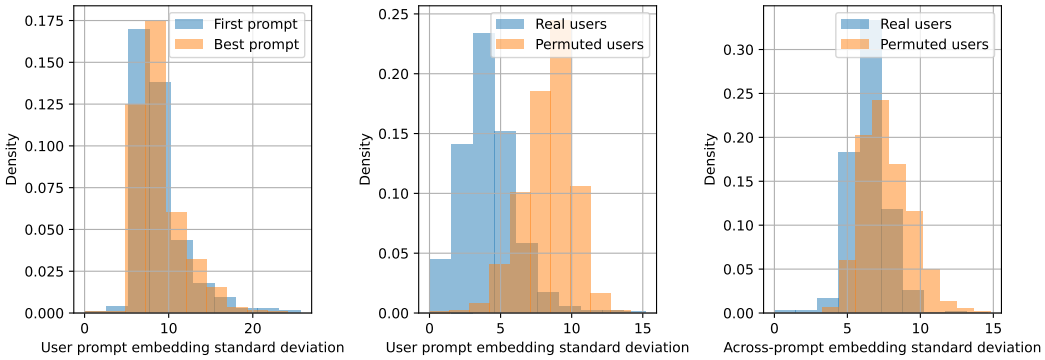

Figure 6. Left: Users submit diverse (across users) prompts, both at beginning and end of interaction. Center: Individual users do not submit diverse prompts. Right: Users have different prompting styles.

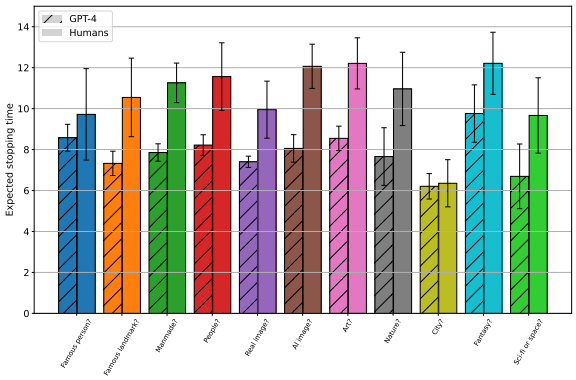

Figure 7. Steerability across image groups for human users and GPT-4 (smaller indicates increased steerability). Bars show average expected stopping time across images in the image group; error bars show standard error.

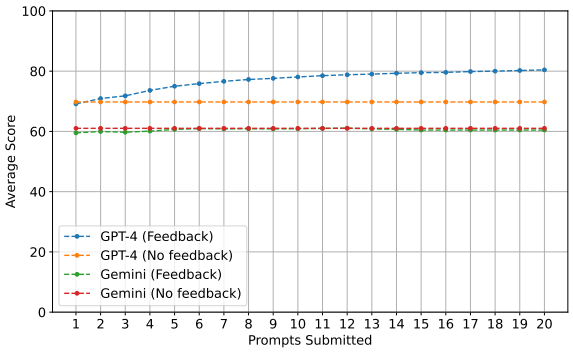

Figure 8. Averaged trajectory of GPT-4 and Gemini. "No feedback" does not use any feedback from previously generated images.

feedback.

When given feedback, GPT-4 consistently improves as it receives more feedback, with minimal improvement after about 15 rounds of feedback. This indicates the model is able to incorporate feedback well. In contrast, Gemini does not improve with feedback. This is expected – the Gemini model we evaluated on was not trained for multi-turn conversations (an API for such a model was not released at the time of writing).

In Figure 7, we compare steerability of the Stable Diffusion model with respect to GPT-4 and humans. Across all tasks, Stable Diffusion is at least as steerable with respect to GPT-4 as it is with humans. This result holds even when we normalize for initial score (i.e., to reduce the effect of baseline prompt writing ability and just examine ability to use feedback), indicating that GPT-4 is better at adapting to feedback compared to human users. This additional plot and more discussion is included in Appendix A.16.

6. Discussion

As demonstrated in our analysis, the *ArtWhisperer* and *ArtWhisperer-Validation* datasets can provide insights into user prompting strategies and enables us to assess model steerability for individual tasks and groups of tasks. What makes our dataset particularly useful is the controlled interactive environment, where users work toward a fixed goal, that we capture data in.

One of the most exciting use cases we see for our dataset is to create synthetic humans for prompt generation. For example, similar to the method described in Promptist (Hao et al., 2022), we imagine fine-tuning a large language model with our dataset to generate prompt trajectories (i.e., rather than an optimized prompt) using similar exploration strategies as a human prompter. These synthetic prompters could be based on multimodal models like OpenFlamingo (Awadalla et al., 2023) or text-only models and use score-feedback to condition the trajectory generation. As an initial proof-of-concept, we fine-tuned a MT0-large model (Muennighoff et al., 2022) model on our dataset and found the fine-tuned model can indeed behave similarly to human users (see

Appendix A.15). It is also feasible that a VLM like the ones assessed in Section 5 could be used as synthetic humans through few-shot prompting, obviating the need for fine-tuning. These synthetic prompters have several uses:

1. Automating measurement of text-to-image model steerability by using synthetic users in place of real human prompters. While we believe our proposed steerability metric is effective, its main limitation currently is the requirement for human annotations.
2. Incorporating steerability in the objective function for text-to-image models. By representing steerability as a function of synthetic users, it becomes possible to explicitly optimize a model for steerability.
3. Generating human readable image captions that are compatible with a Stable Diffusion model by using the synthetic prompter to optimize the token representation of the prompt. This is related to (Zhu et al., 2023), where the authors use models to revise prompts.

Additionally, our dataset can be used for further analysis on human prompting strategies beyond what we discussed in the paper. For example, one question we only touched upon is whether we can compare human prompters to automated prompt optimization methods (e.g., do humans behave similar to some gradient-based optimization approach in the prompt embedding space?). There are also potential uses for crafting better image similarity metrics using the human ratings we collected.

Finally, as discussed in Section 5, our dataset and methodology is useful for assessing vision-language models. In particular, we are able to capture the ability of models to utilize feedback to adapt to a given user or tool.

Impact Statement

Our dataset and findings provide quantitative insights on how people interact with generative AI that we hope can be used to evaluate and design AI that are easier for people to use. There is, of course, a broader concern that bad actors may abuse generative AI models. While we share this concern, addressing it is out of scope for this work.

References

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. Palm 2 technical report, 2023.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P. W., Ilharco, G., Wortsman, M., and Schmidt, L. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Bach, S. H., Sanh, V., Yong, Z.-X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Fevry, T., et al. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*, 2022.

Bard. https://http://bard.google.com, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cascella, M., Montomoli, J., Bellini, V., and Bignami, E. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33, 2023.

Cetinic, E. and She, J. Understanding and creating art with ai: review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–22, 2022.

chatGPT. https://chat.openai.com, 2023.

Chung, J. J. Y., Kim, W., Yoo, K. M., Lee, H., Adar, E., and Chang, M. Talebrush: sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Dakhel, A. M., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M. C., and Jiang, Z. M. J. Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software*, 203:111734, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Hao, Y., Chi, Z., Dong, L., and Wei, F. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ippolito, D., Yuan, A., Coenen, A., and Burnam, S. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*, 2022.

Jahanian, A., Chai, L., and Isola, P. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.

Khandelwal, A., Weihs, L., Mottaghi, R., and Kembhavi, A. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14829–14838, 2022.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

Lexica. https://lexica.art/, 2023.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. A. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35: 1950–1965, 2022.

Liu, V. and Chilton, L. B. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

Midjourney. https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F, 2023.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Nguyen, N. and Nadi, S. An empirical evaluation of github copilot's code suggestions. In *Proceedings of the 19th International Conference on Mining Software Repositories*, pp. 1–5, 2022.

OpenAI. Gpt-4 technical report, 2023.

Oppenlaender, J. Prompt engineering for text-based generative art. *arXiv preprint arXiv:2204.13988*, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Pressman, J. D., Crowson, K., and Contributors, S. C. Simulacra aesthetic captions. Technical Report Version 1.0, Stability AI, 2022. url https://github.com/JD-P/simulacra-aesthetic-captions .

Prolific Academic. https://www.prolific.co, 2023.

Qadir, J. Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1–9. IEEE, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022a.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022b.

Santana, G. Stable-diffusion-prompts. Huggingface Datasets, 2022. URL https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sloan, K. A lawyer used chatgpt to cite bogus cases. what are the ethics? *Reuters*, May 2023. URL https://www.reuters.com/legal/transactional/lawyer-used-chatgpt-cite-bogus-cases-what-ethics-2023-05-30.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

Wu, T., Jiang, E., Donsbach, A., Gray, J., Molina, A., Terry, M., and Cai, C. J. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–10, 2022a.

Wu, T., Terry, M., and Cai, C. J. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2022b.

Wu, X., Sun, K., Zhu, F., Zhao, R., and Li, H. Better aligning text-to-image models with human preference. *ArXiv*, abs/2303.14420, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022a.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022b.

Zhu, W., Wang, X., Lu, Y., Fu, T.-J., Wang, X. E., Eckstein, M., and Wang, W. Y. Collaborative generative ai: Integrating gpt-k for efficient editing in text-to-image generation. *arXiv preprint arXiv:2305.11317*, 2023.

A. Appendix

A.1. Dataset Limitations

A potential limitation in our dataset is the diversity of unique images. However, as shown in our analysis (Sections 3, 4.2), we have multiple illuminating findings despite having <200 unique images in our dataset. In collecting this dataset, there was a tradeoff – given a fixed budget to collect data, we could choose to collect more unique images or collect more user interactions per image. We opted to collect data on a fewer number of unique images with more users interacting with each image, as we believed this data would contain more insights into human interaction with the AI.

A.2. Information on Wikipedia pages scraped

Table 2 presents a list of the Wikipedia pages used to select real-world target images (see Section 2.1). We extracted images from each listed Wikipedia page. We then uniformly sample a category and subsequently sample an image from a page in that category. This ensures a diverse set of images, which is important given that some of the Wikipedia pages contain many more images than others (e.g., `Paris` has 10 times more usable images than `Social_documentary_photography`).

Table 2. Wikipedia images used.

Category	Wikipedia Pages
Art	`Art`, `Fine_art`, `Fine-art_photography`, `History_of_art`, `Painting`
Astro	`Astrophotography`
Buildings	`Architectural_photography`, `Architecture`, `Real_estate`
City	`New_York_City`, `Paris`, `San_Francisco`, `Seoul`
Fashion	`Fashion_design`, `Fashion_photography`, `Model_(person)`
General	`Aerial_photography`, `Culture`, `Documentary_photography`, `Photography`, `Social_documentary_photography`
Landscape	`Landscape_photography`
Nature	`Nature_photography`
Plants	`Flower`
Portrait	`Mug_shot`, `Portrait_photography`, `Selfie`
US	`Americans`, `President_of_the_United_States`, `United_States`
Wildlife	`Aquatic_ecosystem`, `Macro_photography`, `Marine_habitats`, `Wildlife_observation`, `Wildlife_photography`

A.3. Examples of images generated for target parameter selection

Here we provide a few examples of images generated during target parameter selection (see Section 2.1). In Figure 9, we show the original (real) photograph (this is the target image shown to users), some examples of generated images using the caption as a prompt (i.e., the hidden goal for users), and the generated image using the selected parameters with the caption as a prompt (i.e., the generated image when a user finds the "best" prompt using the selected seed, s_{i^*}, from Equation 1). In Figure 10, we show the target image shown to users (t_{i^*} from Equation 2), examples of generated images using the caption on various random seeds, as well as the image generated when using the seed provided to users (i.e., using the seed $s_{i_2^*}$). For both Figures, scores are normalized with respect to the rightmost image in each row – this image is guaranteed to score 100 / 100. Notice that the images generated from other seeds (the center two images) are also similar in quality to the image generated from the selected seed (the rightmost image).

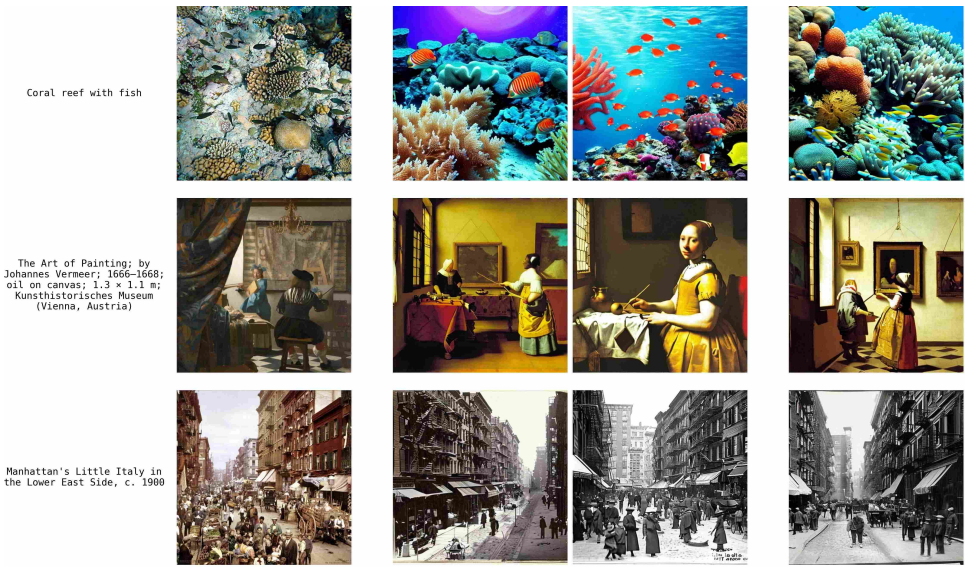

Figure 9. Wikipedia image examples generated using the target prompt (for target selection). From left to right: (1) original Wikipedia captions (the target prompt), (2) original photograph, (3-4) images generated with random seeds, (5) image generated using the fixed seed provided to users.

Figure 10. AI-Generated image examples on the target prompt (for target selection). From left to right: (1) target prompt, (2) target image (uses a different seed than the user's seed), (3-4) images generated with random seeds, (5) image generated using the fixed seed provided to users.

A.4. Scoring function details

In Section 2.2, we defined the scoring function, $score(x, t)$ as

$$score(x,t) = \max(0, \min(100, \alpha \cdot \left|\left| \frac{C(x)}{||C(x)||_2} - \frac{C(t)}{||C(t)||_2} \right|\right|_2 + \beta)),$$

α and β are constants used to scale the embedding distance prior to clipping the score. To select α and β, we used a small dataset of interactions collected by the authors prior to the main dataset collection (this data is not included in the released dataset). This dataset contains groups of images paired with scores in the range $[0, 1]$. For each target image in this small dataset (5 in total), we add the following images to the dataset:

1. AI-generated images that use the target prompt but with a different seed. These images are assigned a score of 1.
2. The images corresponding to the human-generated prompts. These images are assigned a score of 0.5.
3. AI-generated images that use a different prompt than the target prompt. These images are assigned a score of 0.

The intuition here is that with the AI-generated images, we can assume using the same target prompt with a different seed should generate a similar image hence the highest score possible (1). Using a different prompt (from our prompt dataset (Santana, 2022)), however, should result in an entirely different image hence the lowest score possible (0). Images generated by people are assumed to be somewhere in between, hence the score of 0.5.

We then fit a linear regression model to this dataset (to predict score given the CLIP image embedding), using balanced sampling across the image groups. This linear model has parameters

$$\alpha' = -1.503$$
$$\beta' = 1.791$$

Since our score range is $[0, 100]$, we scale the model parameters by 100, resulting in

$$\alpha'' = -150.3$$
$$\beta'' = 179.1$$

We also add a "score adjustment" term that attempts to normalize image difficulty. In particular, for each target image, we compute the un-clipped score for the target image t_k and the image generated using the target prompt with, x_k:

$$unclipped_score(x_k, t_k) = \alpha'' \cdot \left|\left| \frac{C(x_k)}{||C(x_k)||_2} - \frac{C(t_k)}{||C(t_k)||_2} \right|\right|_2 + \beta''$$

This score assess the score a user would receive if they exactly entered the target prompt. We fix this value as 100 (i.e., a perfect score prior to clipping), and set the score adjustment parameter, c_k, to appropriately normalize this score. In particular,

$$c_k = \frac{100}{unclipped_score(x_k, t_k)},$$

and then we obtain target specific parameters,

$$\alpha_k = -150.3 \cdot c_k$$
$$\beta_k = 179.1 \cdot c_k$$

Note that this means each target image may have slightly different parameters as the c_k values vary across images.

A.5. Additional information on running the game

Why we limit user input? We deliberately limited user input to only a prompt, as opposed to giving users access to the random seed or other model hyperparameters. This was done for a few reasons:

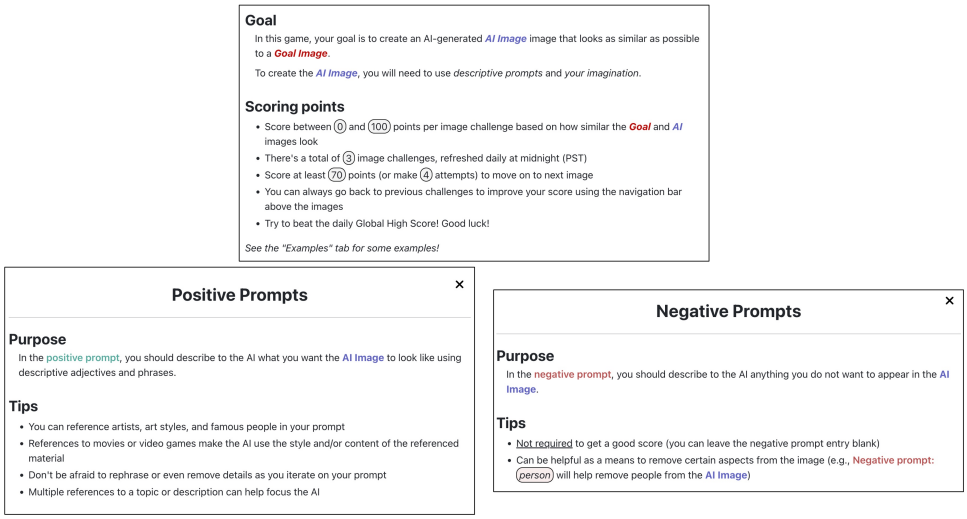

Figure 11. Game instructions

1. We wanted all users who enter the same prompt for a given image to see the same output.

2. We wanted to limit the complexity of the task for users less/unfamiliar with text-to-image models.

3. We wanted users to generate new prompts and not just resample new seeds until getting lucky. While users could still employ a version of this "random resampling" strategy by making small changes to their prompts, we did not want to encourage this practice through a seed parameter. This random resampling strategy, while useful in practice, is not such an interesting result for research purposes as it is easy to simulate random resampling strategies without any user input.

Additional Technical Details For the generative model, we use SD v2.1 (Rombach et al., 2022b) with the DPM Multi-step Scheduler (Lu et al., 2022) and run the model for 20 iterations. AI-generated target images use the same parameters but run for 50 iterations. 20 iterations was selected to limit latency for players to 1-3 seconds depending on the player's internet connection. All images are generated at size 512×512.

Game instructions Game instructions are provided in Figure 11. Here we show the main instructions provided on how to play (top), as well as the tool-tips given for positive prompts (lower left) and negative prompts (lower right).

Crowd workers Crowd workers are adults from the US. They were paid at a rate of $\$12.00$ per hour for roughly 20 minutes of time. Additionally, they were provided bonus payment of between $\$0.10 - \$0.50$ per image they received a perfect score on (depending on the image difficulty). In total, we paid about 600 for recruiting the crowd workers.

A.6. Additional Stats

In Figure 12, we plot the distribution of user scores across target images. The left plot shows the initial and final scores. The shift in score to the right indicates the improvement in generated image similarity to target image. The average initial score is 56.7 (median is 57.0), and the average final score is 73.0 (median is 75.0). On the left, we plot the distribution of score improvement (the difference between the final and initial scores). The large density of 0 or close to 0 improvement is due to the "easiest" target images that users were able to generated similar examples of on their initial attempt (for example, see the first two rows of Figure 19). Most interactions result in a score difference of at most 50 points, as most users score above 50 points on their initial attempt, limiting the amount of improvement possible (the median initial score is 57.0).

In Figure 13, we plot how often a user's score decreases, increases, or remains constant between prompts. Each point represents a series of user interactions to a generate a single target image. We note that the ratio of score increases to score decreases is similar, with the number of score increases being slightly higher. This indicates that score improvement (and so, an increase in the similarity of the generated to target image) may be effectively represented by a stochastic process with a

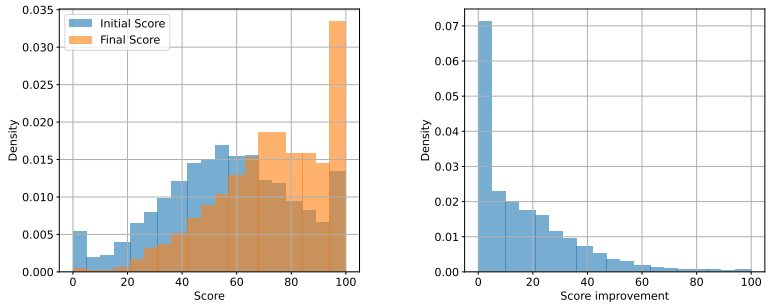

Figure 12. On the left: distribution of initial and final scores (across users and targets). Average initial score is 56.7; average final score is 73.0. On the right: distribution of image improvement (difference between final and initial score). Average score improvement is 16.3.

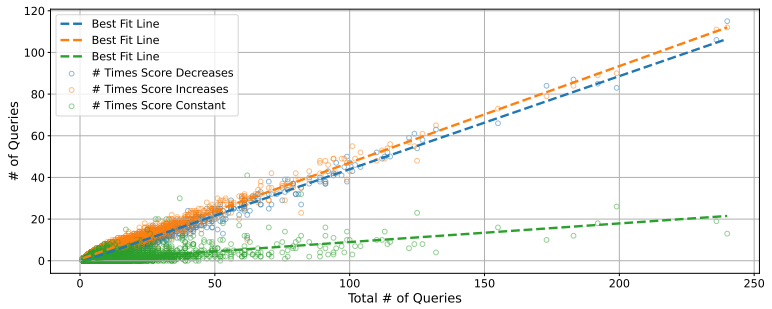

Figure 13. How score changes between queries. Each circle represents a series of user interactions to a generate a single target image. In blue, we plot the number of times a user's score decreases over that interaction (between two queries); similarly, in orange and green we plot the number of times the user's score increases or stays constant between prompts. We also plot best fit lines (fit using linear least squares).

slight bias towards score increases. This observation is the basis on which we construct our proposed steerability metric (i.e., representing score change as a stochastic process).

A.7. ArtWhisperer-Validation Statistics

In Table 3, we provide general statistics about the *ArtWhisperer-Validation* Dataset. This information mirrors that statistics provided in Table 1 for the ArtWhisperer-Validation dataset. We also provide demographic information collected through Prolific in Figure 14. Users were sampled through Prolific to guarantee an even split of male and female users.

A.8. Results for Alternative Text Embeddings

In Figure 15, we replicate the analysis from Figure 6 but using a BERT (Devlin et al., 2018) embedding rather than a CLIP embedding. We note similar findings and significance to the CLIP embedding analysis. We also repeated this analysis using the GLOVE (Pennington et al., 2014) text embedding and again see similar results. This indicates that our analysis is robust

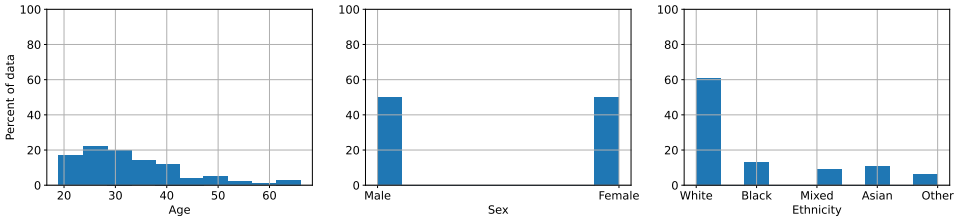

Figure 14. Demographic information for the paid crowd workers in the ArtWhisperer-Validation dataset.

Table 3. ArtWhisperer-Validation Dataset Overview. Each row contains summary data for a different subset of the dataset. Subsets may overlap.

# Players	# Target Images	# Inter-actions	Average # Prompts	Average Score	Median Duration	Category
140	**51**	**4572**	**8.14**	**54.77**	**24 s**	**Total**
73	9	762	7.47	47.55	20 s	Famous person?
134	42	3937	8.19	54.58	23 s	Manmade?
124	26	2154	7.69	54.61	24 s	Real image?
123	24	2424	9.32	50.96	24 s	Art?
94	13	1044	7.05	57.93	25 s	Famous landmark?
72	9	858	8.94	50.33	26 s	Nature?
70	8	730	8.11	60.25	25 s	Sci-fi or space?
94	16	1501	8.25	54.21	24 s	People?
123	25	2418	8.57	54.90	24 s	AI image?
75	9	786	7.42	59.11	23 s	City?
75	9	808	8.16	57.56	26 s	Fantasy?

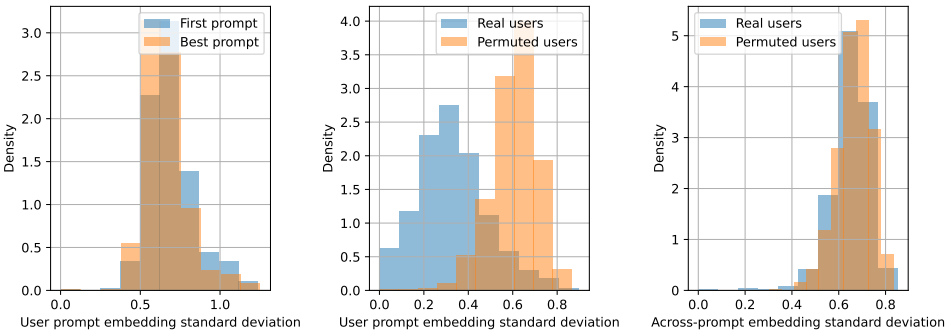

Figure 15. Same plot as Figure 6 but using the BERT (Devlin et al., 2018) text embedding instead of CLIP.

across text representations.

A.9. Additional example images

We provide additional examples of image trajectories and diverse images in Figures 16 and 17.

A.10. Algorithm for steerability

We describe the algorithm for assessing steerability in more detail in Algorithm 1. Here, we define three procedures. `EstimateStoppingTime` estimates the steerability of a target image. We define a set of score bins; we chose 5 equally sized bins so that there was sufficient data to cover each bin. We also use a regularizer, ϵ. What ϵ essentially does is encode a prior that from any given score, the transition to a new score is uniformly random. Then for each target image, we find the empirical score transition probabilities in `EstimateMarkovChain` and use Monte Carlo simulation to estimate the stopping time in `RunMonteCarloEstimation`.

A.11. Steerability across models

In Figure 18, we plot the steerability across SDv2.1 and SDv1.5. As described in Section 4.2, images of nature, sci-fi or space, and real images have the largest differences in steerability between the two models. "Nature" is the only image group with a steerability difference greater than the standard deviation of the mean. Other image groups seem to have similar performance across both SDv2.1 and SDv1.5. This suggests that SDv2.1 only makes minor improvement over SDv1.5 across most image categories.

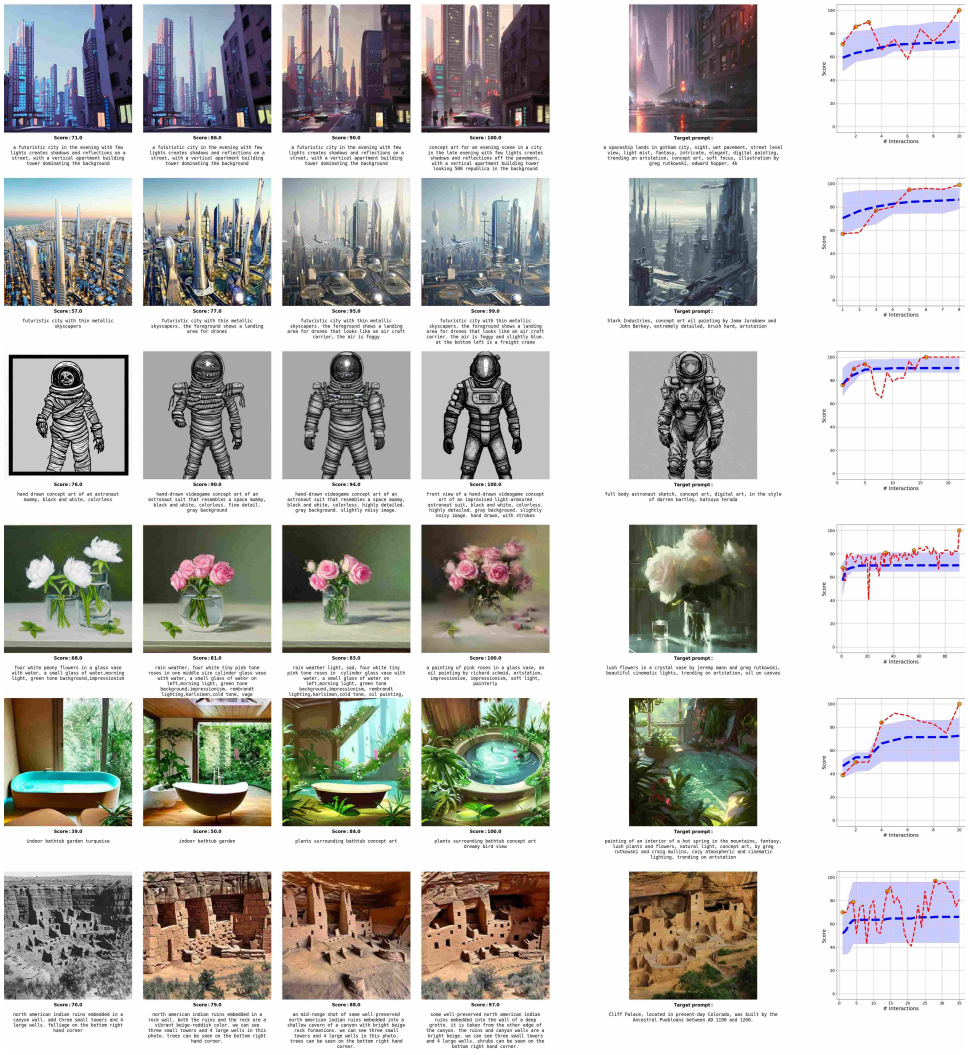

Figure 16. More examples of user trajectories, as in Figure 2.

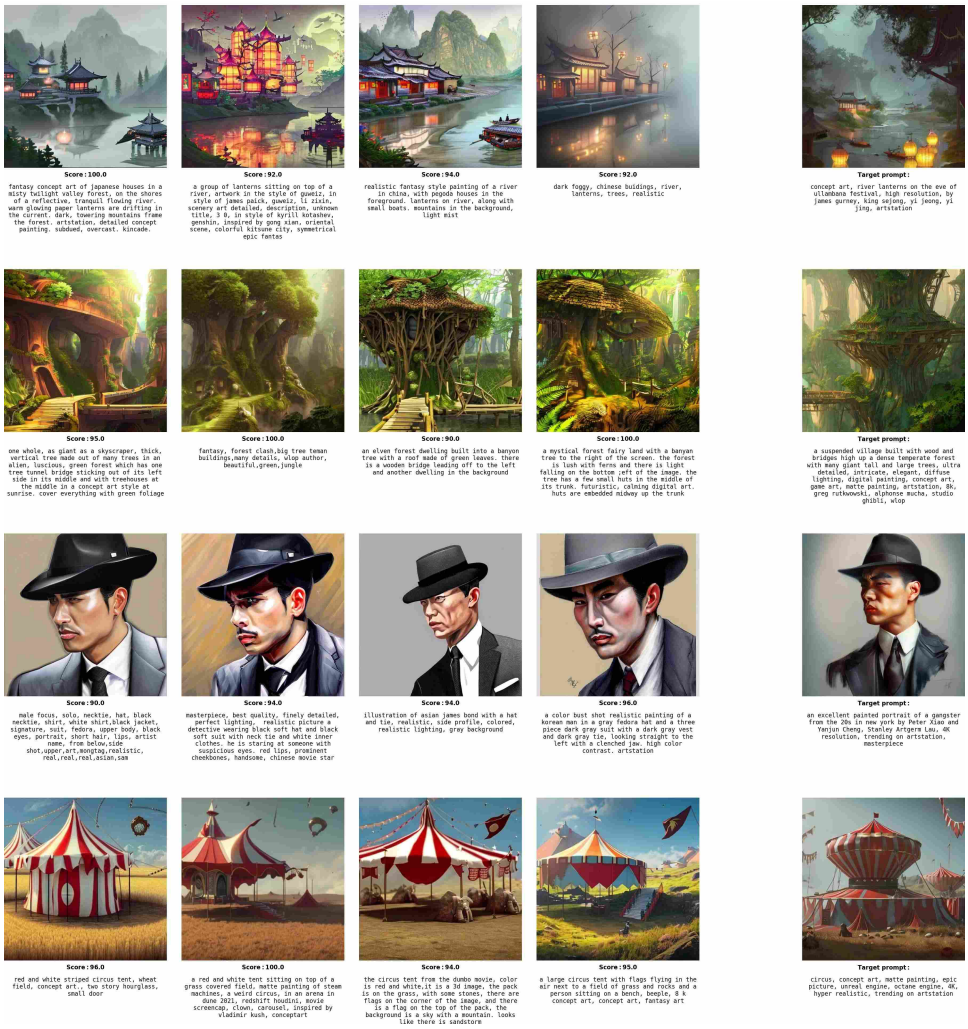

Figure 17. More examples of diverse, high-scoring prompts, as in Figure 4.

Algorithm 1 Image Steerability Estimation

define function `EstimateSteerability`
  Input: set of images, $targetImages$
  Initialize array $steerability$ to track stopping times
  Initialize $scoreBins \leftarrow [[0, 20], [21, 40], [41, 60], [61, 80], [81, 100]]$
  Initialize $\epsilon \leftarrow 1$
  for $image_i \in$ targetImages **do**
    $markov_i \leftarrow$ `EstimateMarkovChain`($image_i$, $scoreBins$, ϵ)
    $steerability[i] \leftarrow$ `RunMonteCarloEstimation`($markov_i$)
  endfor
  Return: $\mathbf{E}[steerability]$

define function `EstimateMarkovChain`
  Input: targetImage, bins, ϵ
  Initialize bin pair count as $counts[(\text{bin}_i, \text{bin}_j)] \leftarrow \epsilon$, using transitions from a dummy node to model the first prompt submitted
  for $\text{user}_i \leftarrow \text{user}_1$ **to** user_n **do**
    for $\text{score}_{i,j} \leftarrow \text{score}_{i,1}$ **to** $\text{score}_{i,d}$ **do**
      Convert $\text{score}_{i,j}$ to bin number, $\text{bin}_{i,j}$
      Increment $counts[(\text{bin}_{i-1,j}, \text{bin}_{i,j})]$ by 1
    endfor
  endfor
  Normalize empirical transition counts to find empirical node transition probabilities
  Define $markov_{\text{targetImage}}$ using the node transition probabilities
  Return: $markov_{\text{targetImage}}$

define function `RunMonteCarloEstimation`
  Input: $markovChain$
  Run Monte Carlo simulation to estimate time to reach the last bin for markovChain, starting from the dummy (initial) node
  Return: *Estimated stopping time*

A.12. Steerability scores for individual images

In Figure 19, we provide some example images along with their steerability scores. Note that more simple images with well-defined content that likely has high presence in the model's training data (e.g., the first two rows–a fly on a leaf; a drawing of Barack Obama, well-known public figure) have smaller steerability values, indicating they are easier to steer. However, more complex content that is also more ambiguous for users (e.g., the last three rows), have larger steerability indicating greater difficulty in steering.

We also find that image steerability is negatively correlated with image variance (across seeds for a fixed prompt). Consider fixing a target prompt and sampling the model across many seeds. When the model outputs a wide variety of images, the variance of the output increases (i.e., computed in the image embedding space). We find that target images that have higher variance are also less steerable (indicated by a higher expected stopping time). In other words, if repeatedly sampling a model with the same prompt can produce a wide range of outputs (high variance), then the model is likely less steerable for the content in those generated images. We plot this result in Figure 20.

A.13. Content knowledge increases steerability

Knowledge of an image's subject matter increases steerability. In Figure 21, we plot the steerability across two groups of individuals – those who had subject matter knowledge and those would did not. To define this split, we only consider the subset of images that contain either a famous person or landmark. Users were defined to have knowledge of the subject matter when they referenced a key word (e.g., the famous person's name) in at least one of their submitted prompts. We then

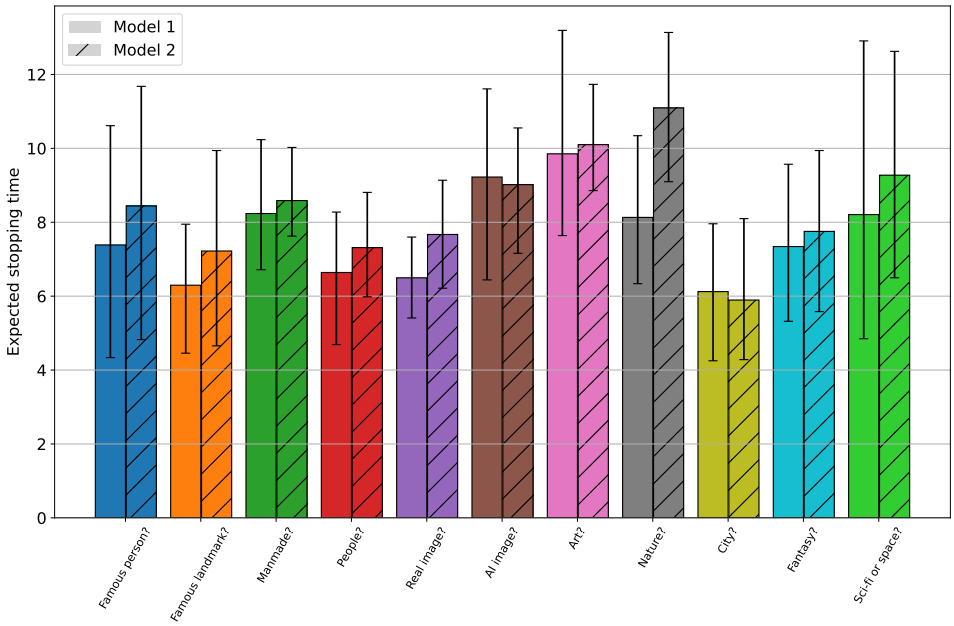

Figure 18. Steerability across image groups, comparing between models using a subset of the images. "Model 1" is Stable Diffusion v2.1 and "Model 2" is Stable Diffusion v1.5. Note only a subset of the images are used to compare the two models, hence the differences between "Model 1" and the results in Figure 7.

computed steerability across these two groups of users. As is indicated in the plot, users with subject matter knowledge found the model to be significantly more steerable.

A.14. Additional discussion around human ratings

Here we provide plots for the analysis using human ratings. In Figure 22, we show a scatter plot of scores and human ratings. We also plot a best fit line which has a correlation of 0.597, indicating that our score function produces values that are indeed similar to the human ratings.

We also provide a plot comparing the steerability value calculated using our score function and calculated using the human ratings in Figure 23. Error bars indicate the standard error. Images depicting sci-fi or space have the greatest difference (humans seem to be harsher judges of their generated images' similarity to the target. However, for most image groups, the two steerability scores are quite close and generally exceed a 95% confidence interval.

A.14.1. On our choice of scoring metric and its relation to human ratings

These results validate our scoring metric based on CLIP. As is well known, embeddings extracted from deep models are reasonable at assessing image similarity (Zhang et al., 2018). We chose to base our scoring metric using the CLIP embedding in particular, however, for two reasons: (1) its recent usage in the literature for effectively representing semantic meaning in images (Radford et al., 2021; Khandelwal et al., 2022) and (2) after testing a number of methods on a small subset of trial data before launching our data collection, we found the CLIP embedding was comparable or better than any of the other methods including ensemble-based approaches. (The methods we tested include embeddings extracted from other deep learning architectures (including ResNet (He et al., 2016) and VGG (Simonyan & Zisserman, 2014)) as well as classical image embeddings (like color histograms).)

Despite this, as is seen in Figure 22, the CLIP-based score does not perfectly correlate with the human ratings. There are two reasons the Pearson correlation coefficient is not larger: (1) noise in the human rating responses and (2) deficiencies in the CLIP embeddings. While improving the image similarity metric is desirable, it was not the focus of our paper. For running ArtWhisperer, we only required a metric that worked reasonably well to generate a "reasonable score" (i.e., which we define as showing at least moderate correlation with recorded human rating assessments). Any subsequent analysis can

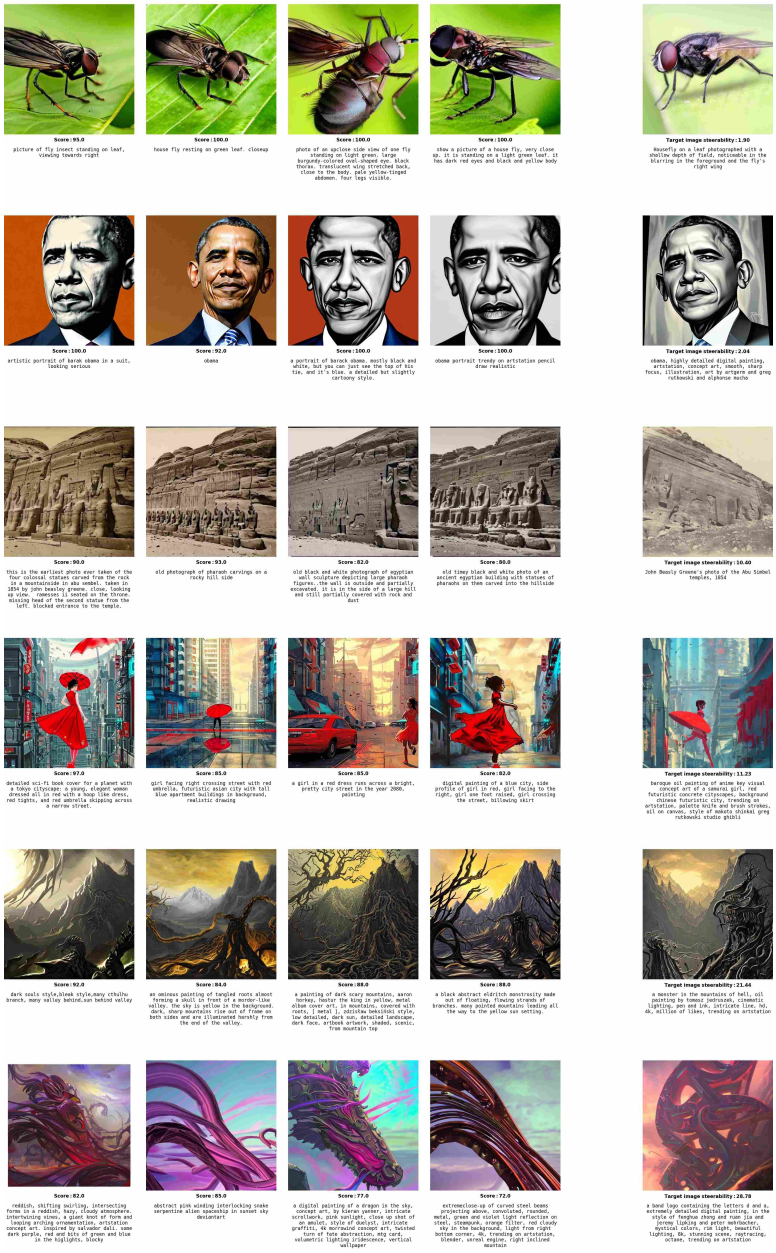

Figure 19. Steerability of individual target images. Each row is a different target image. The first four columns are examples of user submitted prompts with the corresponding AI-generated images. Rightmost column shows the target image along with the *steerability* of that image. A smaller steerability value corresponds to the model being easier to steer for the given target image.

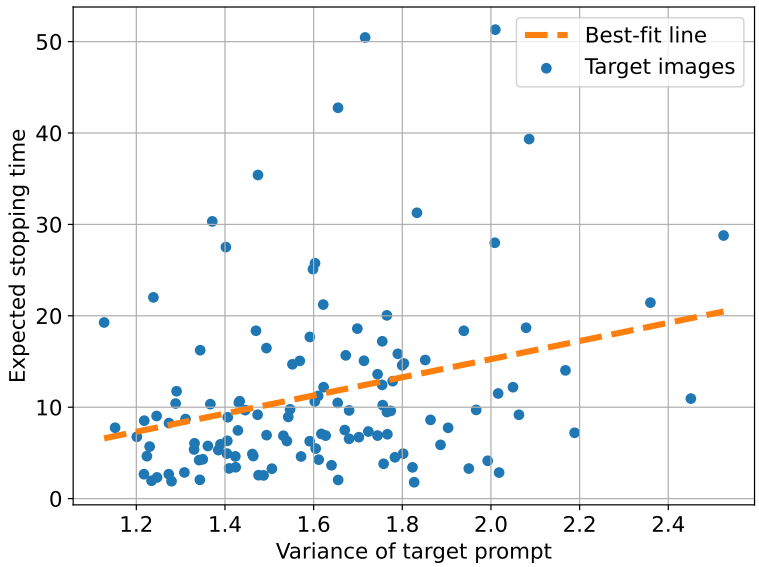

Figure 20. Steerability decreases (expected stopping time increases) when a model has higher variance given a target prompt.

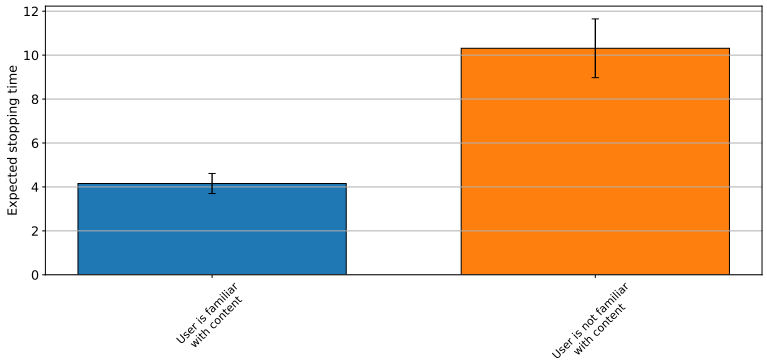

Figure 21. Steerability across familiarity with target image content. When a user is more familiar with the subject matter of the target image, it is significantly easier to steer the model.

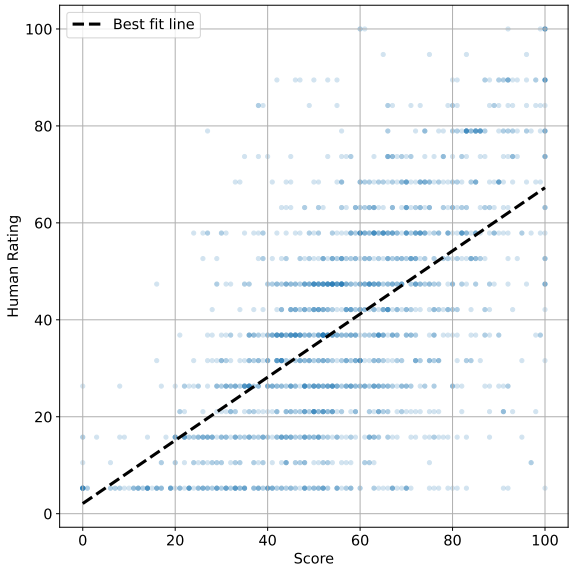

Figure 22. Scatter plot of scores and human ratings. Best fit line (linear regression) is shown in black.

of course use specialized similarity metrics on our data as we release all target and generated images.

The results plotted in Figure 23 also affirm CLIP as a reasonable choice for a metric – the score (based on the CLIP embedding) and human rating, while resulting in different steerability scores, offer similar conclusions across most image categories (e.g., the images that are AI-generated are less steerable than real images).

A.15. Synthetic Prompters

In this section, we describe a proof-of-concept Synthetic prompter based on the ArtWhisperer dataset. We fine-tuned a pretrained MT0-large model (Muennighoff et al., 2022) using the IA3 method (Liu et al., 2022), training for 30 epochs and using a linearly decaying learning rate starting from 10^-3 with the AdamW optimizer (Loshchilov & Hutter, 2017).

Our train dataset is based on the main ArtWhisperer dataset. We first randomly sampled 80% of the ArtWhisperer dataset for training, with sampling taken over the target images (to ensure there are unseen images we can test on). For each user trajectory in the training set, we sampled a sliding window of user prompts. This is done to ensure the total prompt length of the model isn't too long to enable efficient training. Recalling that 50% of the trajectories in the ArtWhisperer dataset have ≤ 5 total prompts, we select a window size of 6 as a reasonable cutoff (allowing a final prompt generation after seeing the 5 previous prompts).

Given that MT0-large is a text-only model, we need a way of encoding the image information for the model to simulate a user prompting on the given image. We opt to simply provide the model with the target prompt; after fine-tuning, the model uses the target prompt to condition its generation but does not repeat it verbatim. In addition to the target prompt, we also provide the model with a history of user prompts and scores indicating how well the image generated by a given prompt matches the target image. For training, we define a loss function on the next-prompt token probabilities to encourage the model to learn to predict subsequent prompts given the prompt and score histories.

The prompt is shown in Figure 24. The "goal prompt" (the target prompt) conditions the model on the target image content. The "user_score_i" values are set based on evaluating the distance between the target and generated image, though in general, we could use any metric of our choosing; the "user_prompt_i" values are set based on previously generated prompts. "user_score_N" is used to condition generation of a new prompt, and enables easy re-sampling by changing the "score" of the new prompt. Note this score is entered pre-evaluation, so we can set it to any value we like; in our evaluations, we set it to a randomly sampled value between 60 and 90.

This approach is similar to prior work like in Promptist (Hao et al., 2022), where the authors fine-tune models to generate high-quality prompts given a human input through supervised learning and reinforcement learning approaches, or in (Zhu

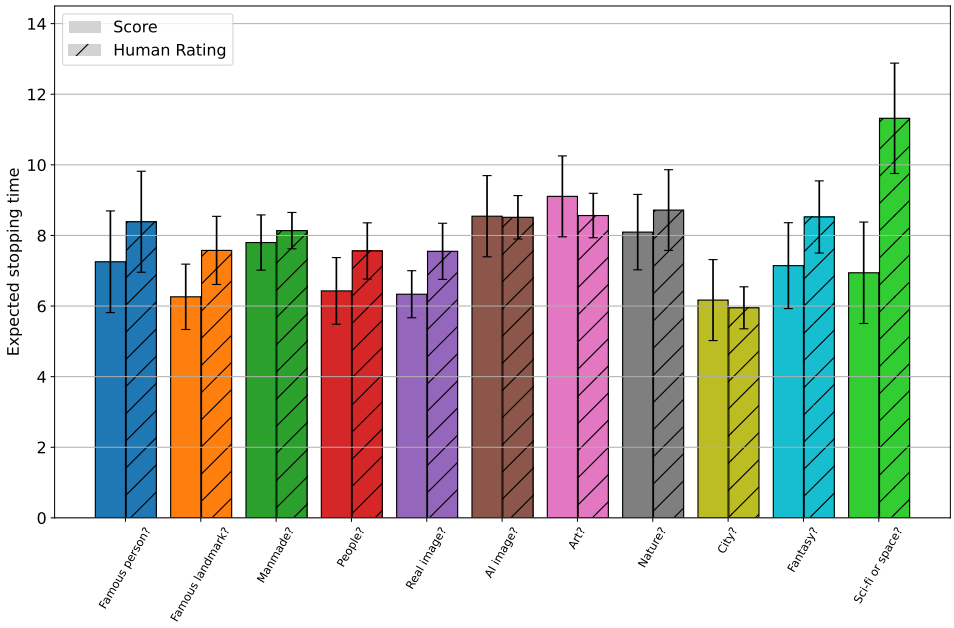

Figure 23. Comparing steerability across image groups when computing steerability using the score function and when using the human rating function.

```
[Goal score]: 100 [Goal prompt]: "{goal_prompt}"
1. [User score]: {user_score_1} [User prompt]: "{user_prompt_1}"
2. [User score]: {user_score_2} [User prompt]: "{user_prompt_2}"
...
N. [User score]: {user_score_N}
```

Figure 24. Prompt format for training and querying the synthetic prompter. $N \leq 6$ in our training and evaluation.

et al., 2023), where the authors also use in-context learning approaches with demonstrations of the initial and final prompts in human trajectories. In both works, the goal is to optimize a prompt. Here, however, we seek to generate a trajectory of prompts that models how a human may behave. Thus while similar in principle, our objective function is different – we train the model to generate the next prompt in a sequence rather than the best performing prompt from an initial prompt.

A.15.1. Analysis of Synthetic Prompters

Here we present an initial analysis that demonstrates the effectiveness of our dataset for training synthetic human prompters. After fine-tuning, we evaluate the synthetic prompter on the held-out test set. The target images in this test set (39 images) were excluded from all training data. For each image, we generate 10 sample trajectories. In Figure 25, we plot the average best score for the synthetic prompter and the real human prompters on the same test data. Note that while the synthetic prompter starts with a higher average, both approach a similar average score after 6 prompts. This suggests that the synthetic prompter indeed shares some similarities with the human prompters.

In Figure 26, we plot two sample trajectories generated by the synthetic prompter. Note that across images, the prompter makes incremental changes; these changes are not restricted to appending phrases, but can involve substitutions and deletions throughout the prompt; this is in contrast to other automated prompters that exclusively complete text (these methods are not designed to simulate human behavior) (Hao et al., 2022).

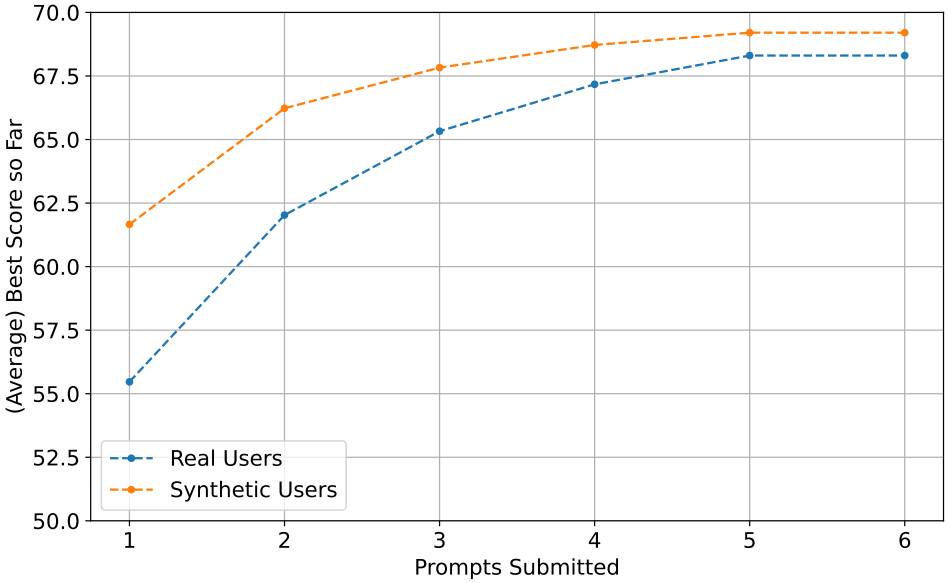

Figure 25. Comparing score progression for the synthetic prompter to real human users.

Figure 26. Example synthetic prompter trajectories. First 3 images show prompt progression for the synthetic prompter. The target image is shown in the rightmost column.

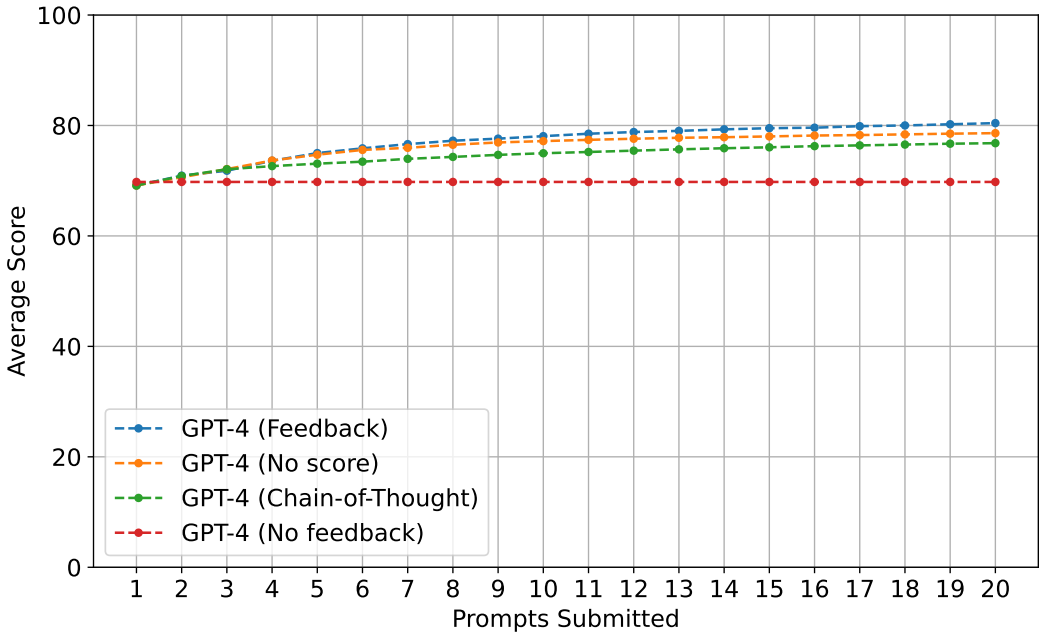

Figure 27. Comparison of additional prompting methods for GPT-4 playing the ArtWhisperer game.

A.16. Assessing VLMs for Feedback Utilization

In Figure 27, we evaluate several methods of prompting. This figure replicates the GPT-4 plots from Figure 8 and adds two additional prompting methods – "No score" and "Chain-of-thought." In "No score," the model is provided feedback as normal, but only the generated image and not the ArtWhisperer score as well. In "Chain-of-thought," the model is provided both the generated image and score feedback, but the VLM prompt used to update the old prompt used for SD uses a chain-of-thought reasoning approach. While all forms of feedback beat the baseline of no feedback, using a score and *not* chain-of-thought reasoning has the highest performance. Interestingly, chain-of-thought reasoning, which typically increases model performance across a variety of tasks (Wei et al., 2022), has lower performance here. Noting that performance for chain-of-thought only begins to diverge after 3 prompts have been submitted, we believe the decrease in performance may be due to the increased context length resulting from chain-of-thought reasoning.

All the prompts used are included in Table 4.

Table 4. Prompts used to have vision language models play ArtWhisperer.

System Prompts	
Purpose	**Prompt**
System prompt (Default)	You are playing a game using a text-to-image AI model. In the game, you will be shown a target image. Your goal is to write a prompt that, when input to the text-to-image model, generates a similar image to the target. You will be scored on the similarity between the generated image and the target image on a scale from 0-100. Your goal is to receive a 100/100. You will first be shown the target image. The game will then proceed in rounds. Each round, you must write a prompt. You will then be shown the image generated by the text-to-image model using your prompt, as well as a score assessing its similarity to the target. You can use this feedback to update your prompt in the next round. Referencing NSFW content will result in a black image and 0/100 score. Be descriptive yet concise. Each prompt must describe the target image in less than 50 words.
System prompt (No score)	You are playing a game using a text-to-image AI model. In the game, you will be shown a target image. Your goal is to write a prompt that, when input to the text-to-image model, generates a similar image to the target. You will first be shown the target image. The game will then proceed in rounds. Each round, you must write a prompt. You will then be shown the image generated by the text-to-image model using your prompt. You can use this feedback to update your prompt in the next round. Referencing NSFW content will result in a black image. Be descriptive yet concise. Each prompt must describe the target image in less than 50 words.

User Prompts	
Purpose	**Prompt**
Initialization prompt (to get the first textual description of the target for all prompting approaches)	Target image: <image_object>
Feedback and prompt updating (Feedback)	Using this prompt, the AI generated the following image: <image_object> You received a score of <ArtWhisperer_score>/100 for this image. Update your prompt to make the generated image closer to the target image.
Feedback and prompt updating (No score)	Using this prompt, the AI generated the following image: <image_object> Update your prompt to make the generated image closer to the target image.
Feedback and prompt updating (Chain-of-Thought)	Using this prompt, the AI generated the following image: <image_object> You received a score of <ArtWhisperer_score>/100 for this image. Update your prompt to make the generated image closer to the target image. Reason step-by-step. First determine what is different between the generated and target images. Then update the prompt to better align the generated image with the target. Delimit the updated prompt with <prompt> tags.