Generation Through Search with Image Aesthetic Assessment

Team 5: Aditya Akula, Justin Blalock, Alex Hobby, David Munechika, Rithvik Rajavelu

ABSTRACT

Text-to-image generative models are capable of generating novel images based on a user-specified prompt and have exploded in popularity recently. Amidst this research, there has been an effort to support the task of "generation through search", essentially allowing users to quickly search for and explore AI-generated images of interest without needing to spend the time and computation to generate the images themselves. However, the existing interfaces do not contain information related to the aethestics of generated images, making it impossible to filter for high-quality images. We propose a system for computing image aesthetic scores on DiffusionDB, the first large-scale prompt gallery dataset, and provide a scalable, interactive interface for users to search for relevant, high-quality images. This work enables users to easily discover and explore images more efficiently and additionally provides insights into effective prompt writing techniques.

1 INTRODUCTION

Text-to-image generative models like Stable Diffusion have garnered significant attention from the generative AI community over the past year. Amidst this research, DiffusionDB was created as the first large-scale prompt gallery dataset for generative models. Furthermore, startups such as Lexica have been created to provide search engines for AI-generated images, so users can quickly find desirable images without needing to perform the slow and expensive computation to generate the image themselves.

However, with both DiffusionDB and current "generation through search" systems, there is no concept of image aesthetics associated with the generated images. This means there is no means for users to filter their search based on the quality of the generated image and thus it may take significantly longer for a user to find an image which is desirable. Furthermore, these search engines are based on textual matches between the search query and the generated image input prompt, and do not allow for exploration of similar, potentially relevant images through the image embeddings themselves.

We propose a new system for computing image aesthetic scores on DiffusionDB images and provide a interactive visualization for users to search for relevant, high-quality images. Formally, our problem definition can be stated as follows: we contribute a novel scalable, interactive visual analytics interface for enabling generation through search capabilities by synthesizing automated image aesthetic assessment scoring with exploration through text and image embedding spaces.

The potential impacts of this project are significant. Filtering the DiffusionDB dataset by image quality would allow for the extraction of a high-quality subset of the data. This could be used to improve the search results so that users can find better looking images faster or it could even be used as a separate dataset itself in order to fine-tune a new diffusion model to produce quality images. Furthermore, having associated IAA scores would enable an analysis on the input prompts to better understand what constitutes an effective prompt (i.e. powerful keywords, overall structure, etc.) so prompt engineers can be more successful. Ultimately, we hope this work will enable users to easily discover and explore images through a more efficient generation through search as well as provide insights into effective prompt writing techniques.

2 PROBLEM DEFINITION

Formally, we define our problem as the extension and synthesization of DiffusionDB with image aesthetic assessment for the purpose of improved generation through search capabilities and enhanced understanding of effective prompt writing techniques. We present our proposed solution as a scalable, interactive visualization interface which supports efficient exploration of relevant, high-quality prompt-image pairs.

3 LITERATURE SURVEY

Diffusion Models. A diffusion model is a type of generative model that learns to denoise inputs by adding and removing noise from sample images. As outlined by Rombach and Blattmann [16], when combined with learning a joint image and text encoding using cross attention, use cases such as Stable Diffusion allow for image generation from text. Findings by Liu and Chilton [11] show that prompt structure highly dictates image aesthetics and work done by Ranford et. al [13] shows that the backbone transformers for cross attention have systematic gaps in understanding. Our work takes into consideration these findings by introducing a novel method to compute image aesthetics and by presenting a visual interface for exploring aesthetic-prompt relationships.

DiffusionDB. DiffusionDB is the first large-scale prompt gallery dataset for text-to-image generative models. While the dataset does include accompanying metadata, it does not contain information about image quality ^{Q2}. We utilize this dataset for a new research direction focused on generation through search with image aesthetic assessment (IAA) [20].

ImageReward extends DiffusionDB by leveraging human annotators to rate image preferences; however, this is limited in that it was only applied to a small subset of DiffusionDB images with no interactive search support. The ImageReward model would be a useful model for our project to test computing IAA scores on images [21]. Similarly, Pick-a-pic [9] created a web app that enables users to generate images and specify their preferences between the results; they then train an aesthetic scoring function that outperforms ImageReward. Since this work has not been applied to DiffusionDB, it can be extended with this project ^{Q2}.

Generation Through Search. Large graph visualizations typically involve displaying a large, interconnected data set through various methods as a way to help with the identification and analysis of patterns. WizMap is an interactive visualization tool that was introduced for exploring large embeddings. The tool scales to millions of embedding points and provides users with an interactive platform that can help analyze the data at varying levels of depth. It has already been used to visualize DiffusionDB prompt embeddings, but is extended in this project to support searching the images themselves [19] ^{Q2}.

The nested Chinese Restaurant Process (*nCRP*) works by categorizing documents into a hierarchical topology and results in a flexible model that can accommodate growing data collections [1]. Semantivisual image hierarchies strategically group and visualize nodes based on commonalities. The model allows for easy clustering of nodes and can be used as a starting point for analysis [10]. These approaches could be used to create a structured index of images to better support search.

Image Aesthetics. Image aesthetic scoring involves judgments of aspects such as lighting, composition, and contrast that contribute to humans' perceptions of image quality [4]. This is particularly applicable to diffusion-based image generation processes, as user-generated images need to have aesthetic qualities, as opposed to simply having accurate features.

Classical approaches to aesthetic scoring largely rely on handcrafted features that evaluate aspects like sharpness, depth, clarity, tone, and colorfulness, as well as composition features that evaluate relative positions of central objects against the background.

Deep approaches use the same general principles, but also make use of techniques like spatial pyramid pooling to capture scene information [2]. More recently, transformer-based models like MUSIQ have also proven to be effective in evaluating quality across different scales [8]. Predictions are then tested against human ratings of images for training. For our project, we use MUSIQ as one measure of image quality.

Prompt Engineering. Prompt engineering involves identifying prompts which can generate the best results for a specific model. It has been shown that methods such as supervised fine-tuning and reinforcement learning can identify more aesthetically pleasing images that still match the original user's intentions [6].

By manipulating embeddings, it is also possible to prompt a diffusion model with an embedding as opposed to a raw text prompt, which could help reduce the negative impact of difficult-to-describe prompts which require trial-and-error prompt engineering [3]. This motivates the need to visualize embeddings and their distance relationships.

PromptMagician is a tool which enables users to explore image results for a prompt to help refine prompts; however, the approach taken is largely natural-language based (TF-IDF) and does not take into account aesthetic aspects of images [5].

4 PROPOSED METHOD

4.1 Intuition

Our proposed method for generation through search should be better than current, state-of-the-art systems because it is the first system to incorporate information about image quality and aesthetics into the search engine. Current systems require more time to identify high-quality, relevant images because there is no mechanism to filter out low quality, generated images. Other systems also lack a means of quickly finding relevant images whereas our system will utilize similarities between image embeddings to cluster similar images together in an interactive visualization interface.

4.2 Algorithms

To compute image aesthetic scores, we utilized the state-of-the-art MUSIQ models from Google Research. Not only do these models perform well for image aesthetic scoring, but they are also specifically designed to accept images of any size and aspect ratio. This is particularly important considering the images in DiffusionDB are not standardized in this regard. There are four open-source pre-trained versions of MUSIQ all trained on different datasets, and we conducted an experiment to determine the best model to use (see Section 5.1).

To support our visualization, we also needed to find a means of clustering together similar images. We use image embeddings, which are lower-dimensional vector representations of images, to accomplish this. To compute embeddings for each image, we used OpenAI's CLIP (Contrastive Language-Image Pre-training) model [15]. CLIP was one of the first multimodal models which is widely used for mapping prompts and images to the same embedding space so it can perform tasks like zero-shot prediction of captions for images [14]. For this project, we specifically use the CLIP image encoder to encode each image as a vector in an embedding space.

4.3 User Interfaces

Our final visualization, shown in Figure 1, has filter options to show the top 500, 2500, 5000, or all data points by IAA score; or all points above an IAA score of 60 or below a score of 15. The user can also sort the results in ascending or descending order of IAA score. To display points, we map the 768-dimensional embeddings of each image into a 2-dimensional space. To do this, we took the embeddings for each image, used the UMAP library for uniform manifold approximation, and generated a 2-dimensional coordinate pair out of each embedding [12]. UMAP is one of the state-of-the-art methods for dimensionality reduction and high-dimensional data visualization which scales better than approaches like t-SNE [18] and PCA [7]. Then, we passed this along with the image and metadata to our web app. By mapping each image embedding to 2-dimensional space, we allow for the visualization of clusters of similar images.

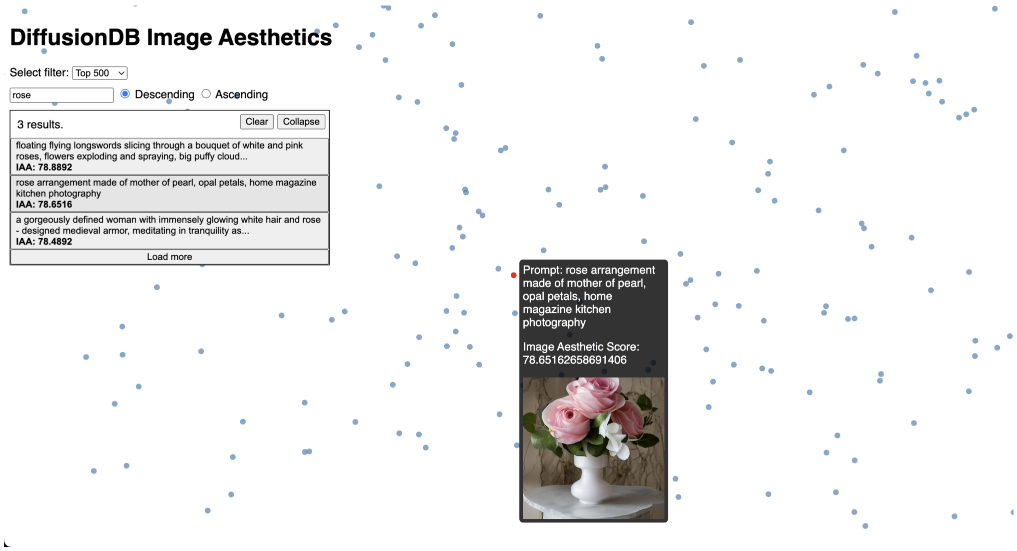
We improve on WizMap by showing both the image and prompt for every single data point in a corresponding tooltip. Image loading is delayed by 200ms to prevent excessive fetches when the user drags their mouse rapidly over a cluster of data points. We also display the aesthetic score for an image and give the user several options for filtering based on aesthetic score so that users can efficiently search and evaluate images based on their specific use case. In addition, our search shows the resultant images along with similar images in the embedding space. We also added a feature to automatically zoom to the data point when a user hovers over a search result. The search results box can also be collapsed / expanded to better view the data points.

4.4 Implementation Details

For this visualization to be useful, it must be quick and reliable for the end-user. To achieve this goal, we have set up a pipeline to allow for efficient computation. Given that there are 2 million images in the DiffusionDB dataset, we needed a way to compute the IAA score for every image in a reasonable timeframe. To do this, we downloaded the images in batches of 1000 then calculated IAA scores for each image using a pre-trained MUSIQ [8] model. We were able to calculate the scores in parallel using Python multiprocessing pools; this process took 10 days to compute 1 million IAA scores, but we ended up only using 100 thousand images in the final interface. Due to the small size of the metadata for each image, we are able to store the image metadata in the browser, but we could scale this by offloading more data to our database to improve loading speed.

We chose to build our visualization using Svelte and D3 similar to WizMap. As opposed to other frameworks such as React, Svelte is compile-based, which enables it to be lightweight and run fast. This allows us to keep our app performant even when we are rendering hundreds of thousands of data points concurrently, a significant concern.

We use Google Cloud Storage to store the images in our visualization. For fast image serving, we use a



DiffusionDB Image Aesthetics

Select filter: Top 500

rose ● Descending ○ Ascending

3 results.

floating flying longswords slicing through a bouquet of white and pink roses, flowers exploding and spraying, big puffy cloud...
IAA: 78.8892

rose arrangement made of mother of pearl, opal petals, home magazine kitchen photography
IAA: 78.6516

a gorgeously defined woman with immensely glowing white hair and rose - designed medieval armor, meditating in tranquility as...
IAA: 78.4892

Prompt: rose arrangement made of mother of pearl, opal petals, home magazine kitchen photography

Image Aesthetic Score: 78.65162658691406

Figure 1: Our current visualization interface clusters similar images based on their CLIP embeddings. Hovering over a node displays the associated image, prompt, and IAA score.

basic CDN to cache images. Since we already have the data loaded in the browser, we use a full-text search package to allow users to search for prompts and/or images locally.

5 EXPERIMENTS & EVALUATION

We conducted a series of experiments and evaluations throughout this project to ensure we selected the best approaches for our system as well as uncover interesting findings as a result. We compiled the following testbed, of which each point is discussed in this section:

1. What is the best approach for automated image aesthetic assessment and how effective is it at capturing human perception??
2. What differences exist between prompts that yield highly aesthetic images and those that do not?
3. How usable, intuitive, and effective is our interactive visualization interface for supporting high-quality, relevant generation through search?

5.1 IAA Model Evaluation

Through a literature search on automated image aesthetic assessment, we discovered that the current state of the art for this task is the MUSIQ architecture. However, various different MUSIQ models have been trained on different datasets, so we performed an experiment to identify the best one for our use case.

To determine what algorithm would be the most viable to use for our project, we decided to test 4 different models that were trained on different datasets but all make use of the MUSIQ algorithm. This experiment was conducted by classifying 100 images from the DiffusionDB dataset as good, bad, or okay. The images were classified based on personal beliefs on the aesthetic quality of each image. We then ran a Python script to allow each model to find the score of an image and then classify them based on where the score fell within a specific range. After conducting this experiment, we were able to determine that there was not a statistically significant difference between the 4 different models. Each of them achieved roughly 85% accuracy at correctly classifying the image into the correct category. Due to there being no significant difference in classification capabilities, we decided to opt for the KonIQ-10k model, which is the default recommended model to use.

5.2 Prompt Aesthetic Analysis

A key innovation of our project is the ability to analyze the effectiveness of prompts in comparison to image aesthetic scores. In other words, users should be able to determine the relationship between various keywords and the resulting IAA score to encourage better prompt engineering. To this end, we identify keywords in prompts and analyze the relation between certain keywords and IAA scores. Similar to previous

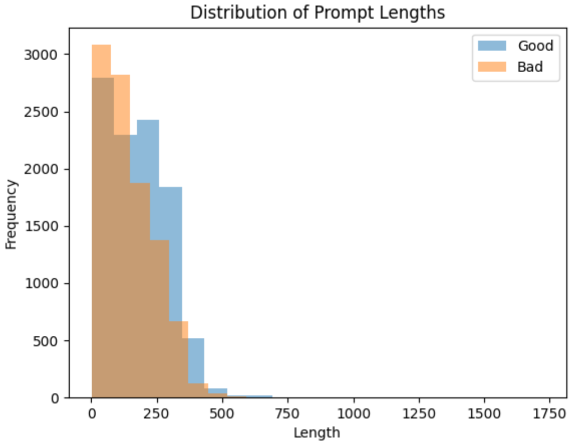


Figure 2: The distribution of prompt lengths for the top 10% and bottom 10%. Good prompts are on average 20% longer than bad prompts.

approaches [5], we calculate keywords using term frequency (tf) and inverse-document frequency (idf):

$$tf_{w,j} = \frac{n_{i,j}}{\sum_k n_{kj}} \tag{1}$$

$$idf_w = log\left(\frac{|D|}{1 + df_w}\right) \tag{2}$$

where $tf_{w,j}$ is the number of times word w occurs in prompt j, $|D|$ is the number of prompts, and df_w is the number of prompts that contain the word w. The idea behind this score is that important words appear frequently in a prompt but a word may not be so important if it occurs frequently in all prompts. We can calculate the tf-idf score as follows:

$$tfidf_{w,j} = tf_{w,j} \times idf_w \tag{3}$$

We compute the tf-idf scores for all words in the separate prompt vocabularies of the top 10% and bottom 10% of images, sorted by IAA score. Figure 3 shows the top 20 most important prompt words for both the low-quality and high-quality images. The color highlights the words which are different between the two lists. Based on these results, we can hypothesize that prompts that guide the generation towards styles such as concept art, illustrations, and digital renders may produce more aesthetic images compared to those with a film, movie, or cinematic style. It is also interesting to note keywords such as *black* appear in the bottom 10% whereas the top 10% contains modifiers such as *sharp*, *focus*, and *intricate* which likely would produce better quality images.

Bottom 10%	**Top 10%**
black: 0.012	render: 0.017
lighting: 0.015	focus: 0.018
light: 0.015	sharp: 0.018
portrait: 0.015	intricate: 0.019
still: 0.015	concept: 0.019
artstation: 0.016	illustration: 0.020
movie: 0.017	highly: 0.022
detailed: 0.018	digital: 0.022
film: 0.019	with: 0.023
from: 0.019	painting: 0.023
with: 0.019	the: 0.024
cinematic: 0.022	on: 0.024
on: 0.023	artstation: 0.026
painting: 0.024	portrait: 0.027
art: 0.026	in: 0.027
in: 0.030	detailed: 0.033
and: 0.038	by: 0.038
the: 0.038	of: 0.039
by: 0.040	art: 0.041
of: 0.042	and: 0.051

Figure 3: Top 20 TF-IDF scores for prompt keywords in the bottom 10% of images and top 10 % of images based on IAA.

We also compare the distribution of prompt lengths between the top 10% and bottom 10% of prompts. We find that the good prompts are on average 20% longer than the bad prompts. This makes sense considering diffusion models tend to benefit from specific details and stylistic modifiers which all contribute to a longer overall prompt length.

5.2.1 Predicting Strong Prompts. To support our goal of assisting users in generation through search and to allow users to understand what resembles a decent prompt, we trained a neural network that takes as input the prompt that the user types in the search bar, and predicts, in real-time, the IAA score of the image that would be generated from that prompt. For the neural network, we took in an input of the vectorized prompt (obtained by assigning an index to each word in the vocab, replacing words with their index, and then padding the vector to be size 100), passed it through an embedding layer, followed by an LSTM, a dense layer, and then output a single IAA score. The choice to use an LSTM was inspired by previous NLP work [17]. Our model achieved a mean absolute error (MAE) score of 7.5449 after 10 epochs. Given the range of our IAA scores (12.1-79.9), we felt that an MAE of 7.5 was too high to give users an accurate understanding of the strength of their prompt, therefore we removed it from the final user interface. Though, being able to predict prompt strength quickly and independent from generation is useful and can be extended with future research.

5.3 User Study

Another experiment we conducted was a user study of our designed system. In the user study, we asked our participants to use the visualization to determine their ideal picture based on a given keyword that will be present in the prompt. After completing this task, the user will then answer a survey that provides responses in a Likert scale of 1-5 with strongly disagree being 1 and strongly agree being 5. In our user study, we were able to determine the average usability and usefulness of our system based on survey results from 7 different participants. From this data, we are able to determine that our tool is easy to use and understand for the average user as we received an average score of 3.71 and 3.86 for these questions. This firmly places the responses on the agreed side. Another important statistic that was measured in our user study was whether or not the user felt that the quality of the pictures increased when the aesthetic score increased. This was our highest scoring area in the user study as we got an average score of 4.29 for this question. Overall, the results from our user study point towards our system being a useful way of determining aesthetic photos quickly and will be easily usable by most users. Below is the average score of all questions asked in the user study:

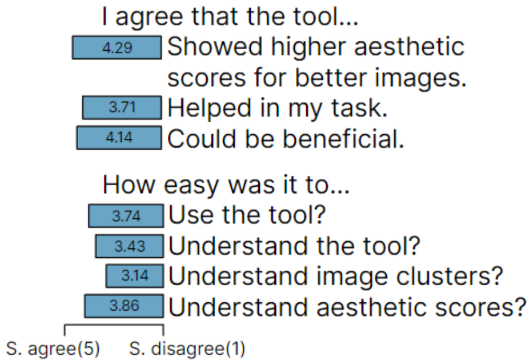


Figure 4: Average response for user study questions

6 CONCLUSION & DISCUSSION

Implications. Based on the results of our evaluation and user study, we have determined that including a dimension for image aesthetics within generation through search helps users efficiently find relevant, high-quality images. The findings of this work also have implications for the overall generative AI community by demonstrating that combining automated IAA scores with generative outputs can yield interesting insights into the quality of the outputs.

Limitations. While our user study demonstrated the effectiveness of our system, it is a prototype that has its limitations. One of the main limitations of the current system is the scale of the visualization. While our system supports 100,000 image-prompt pairs, it would cost a non-trivial amount to host the entire DiffusionDB 2M dataset and introduce additional memory usage pressure on the client side. Furthermore, as the feedback from our user study noted, users felt that there was not much of a guide as to the most effective ways to use the tool. Thus, one way to help our users might be to add a guide or tutorial that would give them a greater understanding of the benefits of the tool. Another important piece of feedback that we received was to potentially add cluster labels. While images that had similar prompts typically grouped and we could clearly see clusters, the lack of a label made it slightly more difficult for users to understand what those clusters specifically represented. A few users also explained that cluster labels could be especially beneficial when making use of our search function.

Future Work. This research supports a number of future research directions that could serve as extensions of this work. One interesting exploration would be to analyze the impact of utilizing a high-scoring IAA dataset for fine-tuning a diffusion model. There have been many recent efforts to fine-tune models on custom datasets based on particular styles or subjects, but it would be fascinating to evaluate whether one could increase the "overall" quality of a diffusion model simply by training it on higher-quality images. Future work could also focus on improving the scalability of the current system and expanding the generation search engine and exploration to include a vastly more comprehensive collection of image-text pairs.

ACKNOWLEDGEMENTS

All team members have contributed a similar amount of effort.

REFERENCES

[1] David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2003. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems* (Whistler, British Columbia, Canada) *(NIPS'03)*. MIT Press, Cambridge, MA, USA, 17–24.

[2] Ying Dai. 2022. Exploring CNN-based models for image's aesthetic score prediction with using ensemble. arXiv:2210.05119 [cs.CV]

[3] Niklas Deckers, Julia Peters, and Martin Potthast. 2023. Manipulating Embeddings of Stable Diffusion Prompts. *arXiv preprint arXiv:2308.12059* (2023).

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (jul 2017), 80–106. https://doi.org/10.1109/msp.2017.2696576

[5] Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. *arXiv preprint arXiv:2307.09036* (2023).

[6] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611* (2022).

[7] Harold Hotelling. 1936. Relations Between Two Sets of Variates. *Biometrika* 28, 3/4 (1936), 321–377. http://www.jstor.org/stable/2333955

[8] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5148–5157.

[9] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569* (2023).

[10] Li-Jia Li, Chong Wang, Yongwhan Lim, David M. Blei, and Li Fei-Fei. 2010. Building and using a semantivisual image hierarchy. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 3336–3343. https://doi.org/10.1109/CVPR.2010.5540027

[11] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 384, 23 pages. https://doi.org/10.1145/3491102.3501825

[12] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML]

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. (2021), 8748–8763.

[14] 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. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[17] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).

[18] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[19] Zijie J. Wang, Fred Hohman, and Duen Horng Chau. 2023. WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings. arXiv:2306.09328 [cs.LG]

[20] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896* (2022).

[21] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977* (2023).