Artist's Assistant: Assist User on Prompt Creation For Text-to-Image Synthesis

Jiayi Pan, Yuanli Zhu, Huanshihong Deng
{jiayipan, leozhu, denghsh}@umich.edu
University of Michigan

1 Introduction

(a) Generated with prompt: "San Francisco".

(b) Generated with prompt: "San Francisco, RPG Reference, Oil Painting, Trending on Artstation, octane render, Insanely Detailed, 8k, HD, unreal 5, DAZ, hyperrealistic, dynamic lighting, intricate detail, summer vibrancy, cinematic".

Figure 1: Two images generated by Stable Diffusion (11) .

Recently, advances in text-to-image synthesis have made it possible to create high-quality, lifelike images using only text descriptions as prompts. However, coming up with the right prompt to create the ideal image can be difficult. For example, while a carefully-engineered prompt (like the one in Figure 1b can produce impressive results, a poorly-written prompt (like the one in Figure 1a) can lead to degradation in image quality. In fact, writing good prompts has been considered a professional skill, known as prompt engineering (4), and high-quality prompts are even being sold on marketplaces [1]. Automating the prompt engineering process has the potential to improve the user experience and make this technology more accessible.

In this project, we make two key contributions:

1. Identify a novel research problem: Given a text description of the scene, our goal is to automatically generate decoration phrases that will make the resulting image more aesthetic while maintaining the meaning.

   Specifically, the input and output of our model is:

   - Input: a textual description of the scene with few or no decorations. For example, "San Francisco" in Figure 1.
   - Output: decoration phrases for that input, which when input together to the text-to-image model, will result in a more aesthetic image while maintaining the semantic of

[1] https://promptbase.com/

image. For example, "RPG Reference, Oil Painting, Trending on Artstation, [...]" in Figure 1b.

For example, in the example of Figure 1, given user input as "San Francisco", we want our model to generate the corresponding decoration phrases like "RPG Reference, Oil Painting..." that will make the resulting synthesized image more aesthetic.

2. Create and evaluate a model: We leverage a Large Language Model for prompt generation and carefully evaluate the performance of our model using both human and machine evaluation methods and our model gets good result with both metrics. The model can be used as a pre-processing techniques for text-to-image algorithms to help researchers or users to create more aesthetic images. The evaluation method used in this project provides others a new way to get the quantified aesthetic score. Besides, the model can assist people to find out the pattern of a good prompt.

There are three main difficulties for this project. Firstly, prompt generation for text-to-image synthesis is a novel direction that has seldom been explored, let alone prompt engineering is itself an emerging topic. Secondly, the evaluation of image synthesis is a lasting hard problem, which we conquered by incorporating both human and machine evaluation. Thirdly, there are no existing dataset containing description-decoration pairs but only raw prompts so we have to preprocess data to train our model.

1.1 Task Allocations

- Jiayi Pan: Project Lead, proposed the project idea and designed the initial version of the system and experiments. He also actively engaged in presentation-making and report writing.
- Yuanli Zhu: Trained the models and did comparison, wrote human evaluation questionnaire and processed the feedback, evaluated the model on LAION aesthetic predictor. He also actively engaged in presentation-making and report writing.
- Huanshihong Deng: Improved the description and decoration spilt method. He also actively engaged in presentation-making and report writing.

2 Related Work

2.1 Text-to-Image Synthesis

Over the past few years, generative models are under fast development in the areas of graphic design, music production, article writing. The field of text-to-image synthesis using deep learning has recently become a hot topic due to diffusion models (1). Diffusion models greatly improves the quality of image generated by GAN, which dominated generation tasks at that time. Now, more and more models, such as Stable Diffusion (11) and DALL-E (9), are released to public and have their own communities for designers or artists sharing their generated image. Figure 1 shows a sample generation result of Stable Diffusion.

2.2 Prompt Engineering

Prompt is the natural language used to specify the task users want to complete and is given to a pre-trained language model to interpret and complete. Prompt engineering is the process of creating a prompt that results in the most effective performance on the downstream task (4). With zero shot capability of foundation models, prompt engineering has arisen across many areas, such as question answering (2), multi-modal learning (13), text-to-image synthesis.

In some areas, the prompt is designed by professionals and fixed for downstream tasks but prompts in text-to-image synthesis will be created purely by end users. Generated image quality can be varied vastly according to the prompt. We can see that figure 1b looks better than 1a by extending the prompt. Therefore, assisting end users in the prompt creation process is an emerging new research field. Recent works include guidelines (5) that help users to write better prompt for text-to-image generative models. Different types of prompt modifiers (7) are classified and verified their utility by comparing the similarity between the output and the ground truth after iterative experimentation. The classification and the proof of necessity for constraints defines the problem more clearly, but it can not work out which prompt may lead to a better output, either. However, using a pre-trained language

model to improve prompting itself in the area of text-to-image synthesis has huge potential to free users from prompt engineering and is largely unexplored. In this work, we aim to make one step to fill this blank and formally propose, to the best of our knowledge, the first such model.

3 Methodology

3.1 Dataset: DiffusionDB

DiffusionDB (15) is a large-scale text-to-image prompt dataset. It contains 14 million prompt-image pairs. For each pair, the image is generated by Stable Diffusion with the corresponding prompt. All data are collected from the official Stable Diffusion Discord server, where users will share their generated images and the corresponding prompts. For our project purpose, we only need to use the prompt part the of dataset.

Though this dataset has large size and is created by Stable Diffusion amateurs for their designing, a problem of this dataset is that there is no guarantee of the quality of the prompts because people who post their result on Discord have different skill levels of writing a prompt. We will show in the later section that the variance of image aesthetics generated by the prompts in this dataset is higher than the one generated by the prompts using our models. However, this dataset is still good enough to train a model to improve a beginner's prompt.

3.2 Preprocessing

As DiffusionDB only contains prompts, we need to preprocess the dataset to create input-output pairs to train our model. The input of our model is a prompt and the output of the model is a list of decorations which can be used to improve the prompt so that Stable Diffusion can generate a better image.

3.2.1 Separating Description and Decorations

We first separate a prompt into a list of clauses by comma and then judge which clause is description and which clause is decoration. The process is shown in figure 2. If the clause has more than 5 words, we directly label it as description since there is only a small chance for a description to be so long.

However, a description can still be short. Our idea is to use part-of-speech tagging and use those tagging sequence to be the input of naive Bayes model, whose output will determine whether the clause is decoration or description. Naive Bayes requires fixed number of features as input and as our input of the model can have at most five taggings as input, we have five part-of-speech taggings as five features of the naive Bayes input. For clause less than 5 words, extra features are set to NULL. For example, a description "a silly goose" will be tagged as [determiner, adjective, noun, null, null] and be the input of naive Bayes model. We labeled 600 clauses manually and train our naive Bayes model. It is able to classify above example as description.

This model is not so accurate. Firstly, naive Bayes treats each word separately and does not consider the context and the whole clause. Secondly, whether a clause is a description or decoration also depends on other clauses in the same prompt and our model treats each clause independently. For example, "long hair" can be a description if user just want an image of long hair but it could a decoration if it co-appears with "a portrait of an astronaut". Thirdly, only using part-of-speech tag may loss some information. However, the processing pipeline is still able to give us a relatively good result.

3.2.2 Generating Input-Output Pairs

Then we delete description-decoration pairs whose number of decorations is less than 6 or no descriptions are identified and randomly select 0-3 decorations from the decoration list and concatenate descriptions to form an input and the rest of the decorations are output. The intuition behind it is that users may write prompts including their own decorations as our model input and our model can predict some other decorations according to user-specified decorations. The deletion of short decorations list is to ensure our model can at least predict 3 decorations even some are selected to be part of the description. Then, we get our input-output pairs. We randomly selected 200k as training dataset, 10k as validation dataset and 10k as testing dataset.

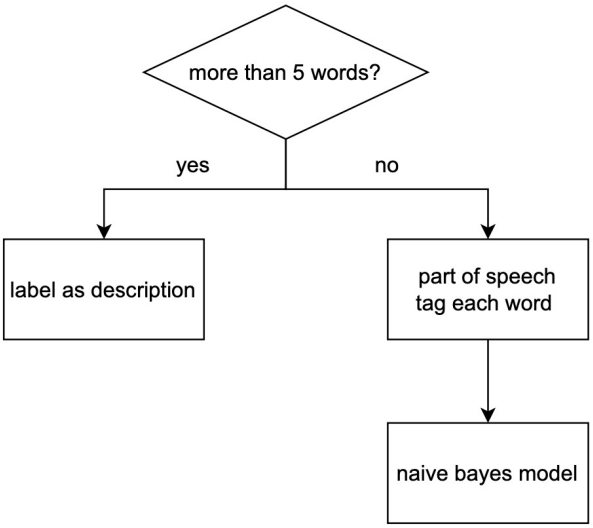


Figure 2: The process of classifying whether a clause is description or decoration.

3.3 Model: BART

Then we use this dataset to finetune BART (3). BART, developed by Facebook, is a transformer (14) model which uses 6 multi-head attention layers in the bidirectional encoder and 6 multi-head attention layers in the left-to-right autoregressive decoder. A larger BART version with 12 layers in each is also provided but that requires more resources to fine-tune so we use the standard version. Encoder takes tokenized sentence as input and output an intermediate representation. Decoder takes the intermediate representation as input and auto-regressively predicts next token and feed the token back to decoder one by one.

BART is pretrained by first corrupting the text with arbitrary noise and then using the corrupted text to predict the original text. This training task enables BART to be trained on unlabeled dataset, which means the dataset can be extremely large. The model and pretraining is shown in figure 3 (3).

Though BART is trained on restoring corrupted text, it also learns text comprehension in this task so it can be particularly effective when fine tuned for other text generation tasks. Though BART itself is trained on an extremely large unlabeled dataset, we are able to use a relatively small labeled dataset to generate a fairly good model.

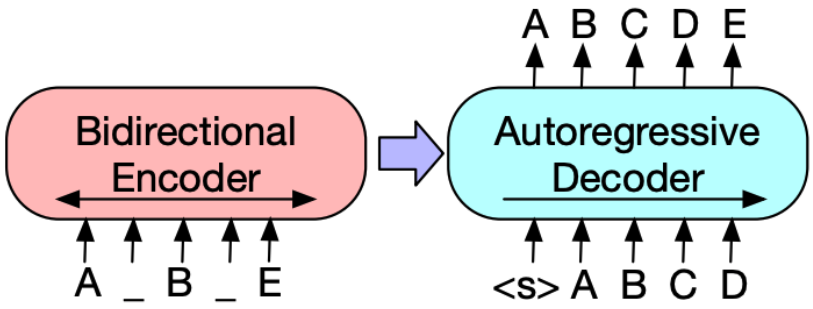


Figure 3: The Encoder takes an corrupted text as an input and output an intermediate representation to decoder. The decoder will receive a start signal <s> and output a token which will then feed back to decoder and repeat this step until decoder output an end signal.

4 Experiments

As our training dataset is DiffusionDB scraped from the Stable Diffusion discord, for all the experiments below, we use Stable Diffusion to generate images.

4.1 Model Training

DiffusionDB (15) has 14 million prompts, which requires too many resources to do the training. Therefore, we randomly chooses 200 thousand data for training, 10 thousand data for validation and 10 thousand data for testing. After pre-processing the data, we get input and output pairs to finetune the BART model. The training parameters are listed in table 1. We trained 5 epochs in total and found that after epoch 3, validation loss hardly drops but training loss drops sharply, which is a sign of overfitting. Since batched inputs are often different lengths, we need to use padding and truncation to create rectangular tensors from batches of varying lengths and we let truncation length be 256.

Table 1: The training parameters of the model.

Parameter	Value
learning rate	5.6×10^{-5}
weight decay	0.01
batch size	8
epoch	3
truncation	256

4.1.1 Recover Original Prompt vs Predict Decorations

As we mentioned above, during training, we use description part to predict decorations instead of recovering the original prompt. In fact, at first, we use the original prompt as the label but get very limited decorations. No matter what the input is, it outputs the descriptions with zero or only a few decorations and their qualities are poor. Here are some examples of our model output with original prompt as labels:

1. input: 'Photo of a man sitting on the car roof in the heavy in front of the city that sank'
   output: 'Photo of a man sitting on the car roof in the heavy in front of the city that'
2. input: 'concept art of a male nobles clothes in the renaissancehand in glove holding laser gun from the side'
   output: 'concept art of a male nobles clothes in the renaissancehand in'
3. input: 'portrait cute cat as csgo characters'
   output: 'portrait cute cat as csgo characters, csgo, csgo'

In fact, we see a lot of results in our test dataset like example 1 and 2, which shows the model is even unable to recover the original descriptions. Therefore, we assumed our model pay too much attention to recovering descriptions as it can make loss decrease fast. That is the reason we turn to only predict decorations, which turns out to be much better.

4.1.2 BART vs T5

We finetune the downstream task on both T5 (10) and BART (3). T5 is another sequence-to-sequence transformer model similar to BART as we introduced. However, the result of T5 version is not satisfiable. The comparison of the three same examples as above is shown below:

1. input: 'Photo of a man sitting on the car roof in the heavy in front of the city that sank'
   T5: 'smashed cars, smashed cars, smashed cars, smashed cars'
   BART: 'highly detailed, digital painting, artstation, concept art, smooth, sharp focus'
2. input: 'concept art of a male nobles clothes in the renaissancehand in glove holding laser gun from the side'
   T5: 'cg animation, riot entertainment, riot entertainment, riot entertainment, riot'
   BART: 'digital painting, concept art, brutal, armory, sad mood,'

3. input: 'portrait cute cat as csgo characters'
   T5: 'renaissance, renaissance, renaiss'
   BART: 'fantasy style, octane render, volumetric lighting, 8k high definition'

Our current assumption for this reason is that BART is trained to recover the corrupted text while T5 is trained on multiple tasks and our task is quite similar to the task which BART is trained on.

4.2 Human Evaluation

As our goal is to help user create more aesthetic images, human evaluation is necessary as aesthetic is quite subjective and hard to measure. We designed a questionnaire using 6 pairs of images generated by Stable Diffusion with one prompt only containing description and another prompt improved by our model. Samples are randomly drawn from the test dataset and there are two questions for each sample, asking them which one is more aesthetic and which one can be well described by the prompt. They do not know in advance which image is generated by the improved prompt. We spread our questionnaire on Piazza and got 16 responses.

From figure 4, we can see in all 6 examples, the image generated by our improved prompt is rated better than the original one. The result shows that our model is able to generate high quality prompt which can make Stable Diffusion generate more aesthetic images.

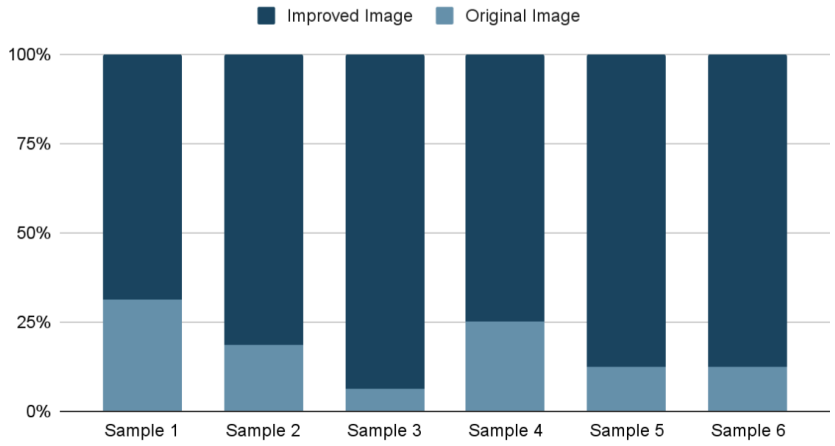


Figure 4: Questionnaire result of the question: which image is more aesthetic.

Besides aesthetics, another thing we care is whether our model changes the meaning of the original prompt. Description is the content users want in their image, and if we add too much distraction in the prompt, Stable Diffusion may not be able to identify what users want to convey and focuses too much on decorations, which is absolutely not we want. From figure 5, we can see that in all six samples, most people think neither or both of images can be well described by the model. The portion of thinking the improved image or the original image is closer to the prompt is about the same. Therefore, no observable change of the prompt meaning is found.

4.3 CLIP Aesthetic Score

Though human evaluation is quite suitable as an aesthetic metric, it can be hardly done in large scale. Therefore, besides using human evaluation, we also explore using a machine-based aesthetic metric, LAION aesthetics predictor (12), to evaluate our model. LAION aesthetics predictor is a simple neural net on top of CLIP (8) to predict the aesthetic score. CLIP is a very popular models nowadays. It jointly trains a text encoder like transformer and an image encoder like CNN to predict the correct pairings of a batch of input. In this way, both text encoder and image encoder are able to generate high quality text embedding and image embedding. LAION aesthetics predictor freezes the weight of the image encoder and directly takes the CLIP image embedding as input and output aesthetic score. The model is trained on SAC (Simulacra Aesthetic Captions), which is a dataset with 176000 image-rating pairs. The rating ranges from 1 to 10.

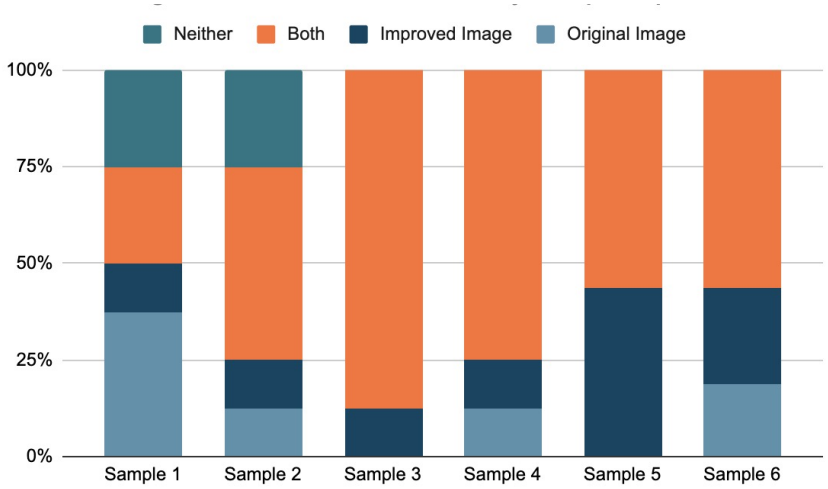


Figure 5: Questionnaire result of the question: which image can be well described by the prompt.

Table 2: Mean and variance of rating 1000 samples with LAION aesthetic predictor.

	mean	variance
raw	5.802	0.291
model	5.981	0.227
human	6.032	0.234

We randomly selected 1000 samples from the testing dataset. For each sample, we let Stable Diffusion generate images using raw description, description with model-generated decorations, description with human-labeled decorations. Then let LAION aesthetics predictor rate those images.

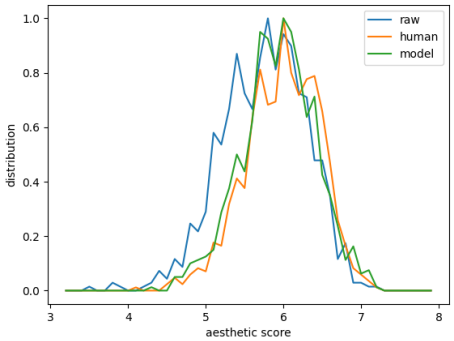
Figure 6a shows the distribution of the score and figure 6b shows the cumulative distribution, which is more clear. Table 2 shows the mean and variance. We can see that the mean score of the image improved by our model-predicted decorations is obviously higher than the mean score of image generated by raw description and the variance is lower. In addition, both mean and variance is very closed to image improved by human-labeled decorations. Here, the variance of human-labeled decorations is slightly higher than the variance of model-generated decorations and that may be shown that the quality of decorations labeled by different people varies. Therefore, it can be shown that our model has the ability to improve prompt to generate more aesthetic images and it can be done more stably.

4.4 Case Study

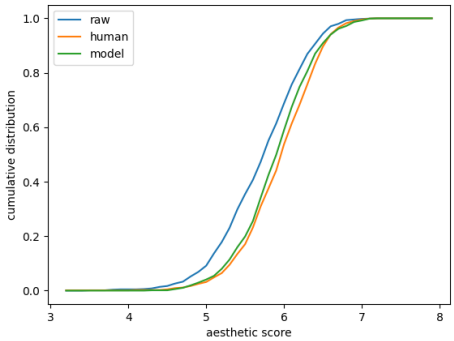
It is shown above that our model can improve prompt to make Stable Diffusion generate more aesthetic image. In this section, two cases are analyzed to see what decorations are added and why adding them can make the image look better. Below two examples are picked from human evaluation questionnaire and the image generated by our improved prompt are thought to be more aesthetic by most people.

The first example is shown figure 7. We can see that our model is able to add some concrete details like "epic sky" are "low angle". Those details can be easily neglected by a beginner and did not change the original meaning of the prompt but they are essential for Stable Diffusion to generate an impressive image. In addition, it also adds style-related decorations like "concept art", which shows our model is able to choose the styles that fits what prompt describes.

Another interesting example is shown in figure 8. The model adds some artists' names as decorations like Tristan Eaton and Victo Ngai. Figure 9 shows an example of their work. We can see that our model is able to choose artists whose painting topic and style fitting the prompt and Stable Diffusion is able to transfer their styles to the generated image.



(a) The CLIP aesthetic score distribution of image generated by raw description, description with model-predicted decorations, description with human-labeled decorations.



(b) The CLIP aesthetic score cumulative distribution of image generated by raw description, description with model-predicted decorations, description with human-labeled decorations.

Figure 6: Experiment result of rating 1000 samples with LAION aesthetic predictor.

(a) Image generated by the original prompt.

(b) Image generated by the enhanced prompt.

Figure 7: Two images generated by Stable Diffusion. Figure 7a is generated by the original prompt: "a levitating kingdom floating above the ground". Figure 7b is generated by the prompt improved by our model: "a levitating kingdom floating above the ground, cinematic view, epic sky, detailed, concept art, low angle".

Those generated results actually corresponds to the current findings which believes that adding more details and specifying styles can help text-to-image models to generate better images(6).

(a) Image generated by the original prompt.

(b) Image generated by the enhanced prompt.

Figure 8: Two images generated by Stable Diffusion. Figure 8a is generated by the original prompt: "a lot of flowers and wires on body". Figure 8b is generated by the prompt improved by our model: "a lot of flowers and wires on body, tristan eaton, victo ngai, artgerm, rhads".

(a) Art work by Trison Eaton.

(b) Art work by Victo Ngai.

Figure 9: The work of artist who is chosen as the decorations by the model in example 8.

References

[1] DHARIWAL, P., AND NICHOL, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems* (2021), M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., pp. 8780–8794.

[2] KHASHABI, D., MIN, S., KHOT, T., SABHARWAL, A., TAFJORD, O., CLARK, P., AND HAJISHIRZI, H. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online, Nov. 2020), Association for Computational Linguistics, pp. 1896–1907.

[3] LEWIS, M., LIU, Y., GOYAL, N., GHAZVININEJAD, M., MOHAMED, A., LEVY, O., STOYANOV, V., AND ZETTLEMOYER, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL* (2020).

[4] LIU, P., YUAN, W., FU, J., JIANG, Z., HAYASHI, H., AND NEUBIG, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* (aug 2022).

[5] LIU, V., AND CHILTON, L. B. Design guidelines for prompt engineering text-to-image generative models, 2021.

[6] LIU, V., AND CHILTON, L. B. Design guidelines for prompt engineering text-to-image generative models, 2021.

[7] OPPENLAENDER, J. A taxonomy of prompt modifiers for text-to-image generation, 2022.

[8] RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., KRUEGER, G., AND SUTSKEVER, I. Learning transferable visual models from natural language supervision, 2021.

[9] RAMESH, A., DHARIWAL, P., NICHOL, A., CHU, C., AND CHEN, M. Hierarchical text-conditional image generation with clip latents, 2022.

[10] ROBERTS, A., AND RAFFEL, C. Exploring transfer learning with t5: the text-to-text transfer transformer. *Accessed on* (2020), 23–07.

[11] 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)* (June 2022), pp. 10684–10695.

[12] SCHUHMANN, C., BEAUMONT, R., VENCU, R., GORDON, C., WIGHTMAN, R., CHERTI, M., COOMBES, T., KATTA, A., MULLIS, C., WORTSMAN, M., ET AL. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

[13] TSIMPOUKELLI, M., MENICK, J., CABI, S., ESLAMI, S. M. A., VINYALS, O., AND HILL, F. Multimodal few-shot learning with frozen language models, 2021.

[14] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U., AND POLOSUKHIN, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc.

[15] 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:2210.14896 [cs]* (2022).