Best Prompts for Text-to-Image Models and How to Find Them

Nikita Pavlichenko*
Toloka
Belgrade, Serbia
pavlichenko@toloka.ai

Dmitry Ustalov
Toloka
Belgrade, Serbia
dustalov@toloka.ai

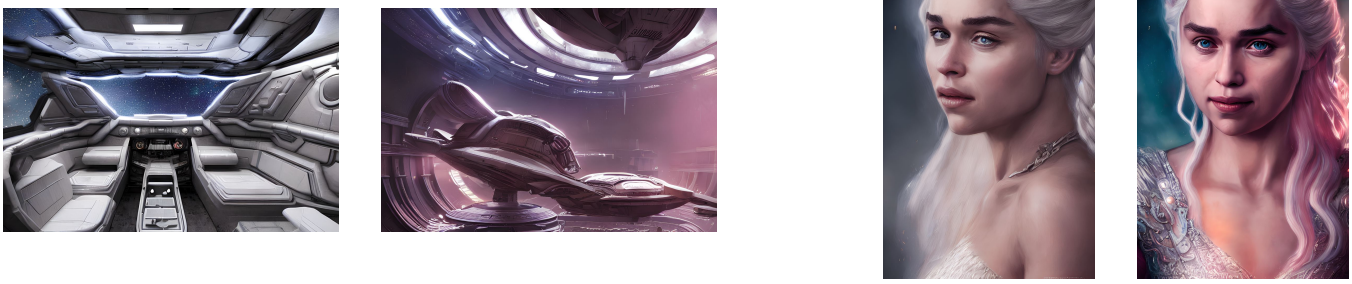

Figure 1: Comparison of Stable Diffusion model-generated images using popular and custom keywords. Two pairs of images are shown, with each pair consisting of an image generated with the top-15 most popular keywords (left) and an image generated with keywords found by our method (right). The left pair of images depicts the "interior of an alien spaceship," while the right pair depicts "daenerys targaryen queen." Descriptions are cherry-picked.

ABSTRACT

Advancements in text-guided diffusion models have allowed for the creation of visually appealing images similar to those created by professional artists. The effectiveness of these models depends on the composition of the textual description, known as the *prompt*, and its accompanying keywords. Evaluating aesthetics computationally is difficult, so human input is necessary to determine the ideal prompt formulation and keyword combination. In this study, we propose a human-in-the-loop method for discovering the most effective combination of prompt keywords using a genetic algorithm. Our approach demonstrates how this can lead to an improvement in the visual appeal of images generated from the same description.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in visualization*; • **Information systems** → **Query reformulation**; Image search.

KEYWORDS

human feedback, text-to-image generation, genetic optimization, aesthetics evaluation

*Primary contact person.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592000

ACM Reference Format:
Nikita Pavlichenko and Dmitry Ustalov. 2023. Best Prompts for Text-to-Image Models and How to Find Them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592000

1 INTRODUCTION

Recent progress in computer vision and natural language processing has enabled a wide range of possible applications to generative models. One of the most promising applications is text-guided image generation (text-to-image models). Solutions like DALL-E 2 [14] and Stable Diffusion [16] use the recent advances in joint image and text embedding learning (CLIP [13]) and diffusion models [19] to produce photo-realistic and aesthetically-appealing images based on a textual description.

However, in order to ensure the high quality of generated images, these models need a proper *prompt engineering* [7] to specify the exact result expected from the generative model. In particular, it became a common practice to add special phrases (*keywords*) before or after the image description, such as "trending on artstation," "highly detailed," etc. Developing such prompts requires human intuition, and the resulting prompts often look arbitrary. Another problem is the lack of evaluation tools, so practically, it means that the user subjective judges the quality of a prompt by a single generation or on a single task. Also, there is currently no available analysis on how different keywords affect the final quality of generations and which ones allow to achieve the best images aesthetically.

In this work, we want to bridge this gap by proposing an approach for a large-scale human evaluation of prompt templates

using crowd workers. We apply our method to find a set of keywords for Stable Diffusion that produces the most aesthetically appealing images. Our contributions can be summarized as follows:

- We introduce a method for evaluating the quality of generations produced by different prompt templates.
- We propose a set of keywords for Stable Diffusion v1.4 and experimentally show that it improves the aesthetics of the generated images.
- We release all the data and code that allow to reproduce our results and build solutions on top of them, such as finding even better keywords and finding them for other models.

2 PROMPTS AND HOW TO EVALUATE THEM

Consider a standard setup for generative models with text inputs. A model gets as an input a natural language text called *prompt* and outputs a text completion in the case of the text-to-text generation or an image in the case of text-to-image generation. Since specifying the additional information increases the quality of the output images [7], it is common to put specific keywords before and after the image description:

$$\text{prompt} = [\text{kw}_1, \ldots, \text{kw}_{m-1}]\ [\text{description}]\ [\text{kw}_m, \ldots, \text{kw}_n].$$

Consider a real-world example when a user wants to generate an image of a cat using a text-to-image model.[1] Instead of passing a straightforward prompt *a cat*, they use a specific prompt template, such as *Highly detailed painting of a calico cat, cinematic lighting, dramatic atmosphere, by dustin nguyen, akihiko yoshida, greg tocchini, greg rutkowski, cliff chiang, 4k resolution, luminous grassy background*. In this example, the **description** is *painting of a calico cat* and the **keywords** are *highly detailed, cinematic lighting, dramatic atmosphere, by dustin nguyen, akihiko yoshida, greg tocchini, greg rutkowski, cliff chiang, 4k resolution, luminous grassy background*.

Since aesthetics are difficult to evaluate computationally, we propose a human-in-the-loop method for evaluating the keyword sets. Our method takes as an input a set of textual image descriptions $\mathcal{D}$, a set of all possible keywords $\mathcal{K}$, and a set of the keyword set candidates $\mathcal{S}$ and outputs a list of keyword sets $s_i \subseteq \mathcal{K}, s_i \in \mathcal{S}$ in the increasing order of their aesthetic appeal to humans. Since it is challenging for annotators to directly assign scores for images or rank them, we run pairwise comparisons of images generated from a single description but with different keyword sets and then infer the ranking from the annotation results. Our algorithm can be described as follows:

(1) For each pair of a description $d_i \in \mathcal{D}$ and a keyword set $s_j \in \mathcal{S}$, generate four images $I_{ij} = \{I_{ij_1}, \ldots, I_{ij_4}\}$.
(2) For each image description $d_i \in \mathcal{D}$, sample $nk \log_2(n)$ pairs of images (I_{ij}, I_{ik}) generated with different keyword sets, where n is the number of keyword sets to compare, and k is the number of redundant comparisons to get the sufficient number of comparisons [9].
(3) Run a pairwise comparison crowdsourcing task in which the workers are provided with a description and a pair of images, and they have to select the best image without knowing the keyword set.
(4) For each description $d_i \in \mathcal{D}$, aggregate the pairwise comparisons using the Bradley-Terry probabilistic ranking algorithm [1], recovering a list $r_i = s_1 \prec \cdots \prec s_n$ of keyword sets ordered by their visual appeal to humans.
(5) For each keyword set, compute the average rank in the lists recovered for the descriptions.

As a result, we quantify the quality of a keyword set as its rank averaged per description.

3 ITERATIVE ESTIMATION OF THE BEST KEYWORD SET

One of the advantages of our approach is that the keywords can be evaluated iteratively. Once we have compared a number of keyword sets, we can request a small additional number of comparisons to evaluate the new set of keywords. This allows us to apply discrete optimization algorithms, such as a genetic algorithm, to retrieve from a large pool of keywords the most influential keywords.

Figure 2 represents a scheme of our approach. We pick a set of keyword sets for initialization, rank the keywords using the approach in Section 2, and use it as an initial population for the genetic algorithm. Then we repeat the following steps multiple times to obtain the best-performing keyword set.

(1) Obtain the next candidate keyword set s_j based on quality metrics of currently evaluated keyword sets using the genetic algorithm. We present the details on a particular variation of a genetic algorithm we use in Section 4.1.
(2) For each image description $d_i \in \mathcal{D}$, sample $k\left((n+1)\log_2(n+1) - n\log_2 n\right)$ pairs (I_{ik}, I_{ij}) of images generated using keywords from the new candidate set and already evaluated keyword sets. We do this to sustain $kn\log_2 n$ comparisons in total.
(3) Evaluate the quality of the obtained keyword set (steps 3-5 in Section 2).

4 EXPERIMENT

We perform an empirical evaluation of the proposed prompt keyword optimization approach in a realistic scenario using the publicly available datasets.

4.1 Setup

To construct a set of possible keywords, we have parsed the Stable Diffusion Discord[2] and took the 100 most popular keywords. For image descriptions, we decided to choose prompts from six categories: *portraits*, *landscapes*, *buildings*, *interiors*, *animals*, and *other*. We took twelve prompts for each category from Reddit and https://lexica.art/ and manually filtered them to obtain only raw descriptions without any keywords.

We use a simple genetic algorithm to find the optimal prompt keyword set. The algorithm was initialized with two keyword sets: one is an empty set, and another set contained the 15 most popular keywords that we retrieved before. We limited the maximum number of output keywords by 15 as otherwise, the resulting prompts became too long.

[1] https://lexica.art/?q=a+cat&prompt=28f5c644-9310-4870-949b-38281328ffd0

[2] https://discord.com/invite/stablediffusion

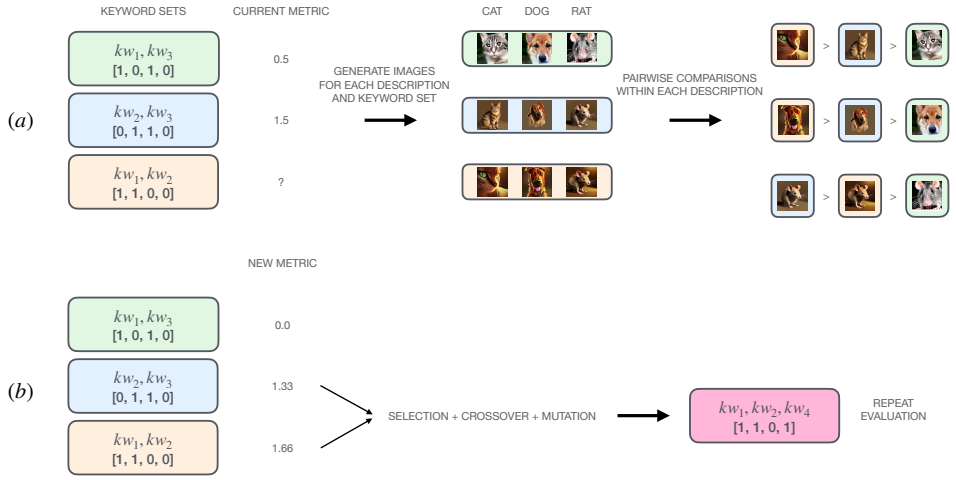

Figure 2: A scheme of genetic optimization of keyword sets. (a) Evaluation of a new candidate keyword set: first, we generate images for all descriptions with a new keyword set; second, we run pairwise comparisons of generated images within each description between the previous and new keyword sets to obtain the ranking. The average rank of keyword sets is a quality metric. (b) We take two keyword sets with the highest rank and perform crossover and mutation to obtain a new candidate, which is then evaluated according to scheme (a). The process is repeated for the pre-determined number of iterations.

Interior of an alien spaceship	
Image L1	Image R1
Image L2	Image R2
Image L3	Image R3
Image L4	Image R4
Which set is better?	☐ Left ☐ Right

Figure 3: Textual pseudographics of the annotation interface. A crowd worker sees two sets of four images generated for a single description but with different keyword sets (one on the left and one on the right) and needs to choose the more aesthetically-pleasing set of images.

In order to evaluate the keyword sets, we generate four images for each prompt constructed by appending comma-separated keywords to the image description in alphabetical order. Each image was generated with the Stable Diffusion model [16] with 50 diffusion steps and 7.5 classifier-free guidance scale using the DDIM scheduler [20]. Then, we run crowdsourcing annotation on the Toloka crowdsourcing platform.[3] The crowd workers need to choose the most aesthetically-pleasing generated images in $3n \log_2 n$ pairs (we set $k = 3$ as we have a limited budget) for each image description, where n is the number of currently tried keyword sets. Textual pseudographics of the annotation interface is shown in Figure 3.

Since crowdsourcing tasks require careful quality control and our task involved gathering subjective opinions of humans, we followed the synthetic golden task production strategy proposed for the IMDB-WIKI-SbS dataset [11]. We randomly added comparisons against the images produced by a simpler model, DALL-E Mini [4]. We assumed that DALL-E Mini images are less appealing than the ones generated by Stable Diffusion, and choosing them was a mistake. Hence, we suspended the workers who demonstrated an accuracy lower than 80% on these synthetic golden tasks.

After the annotation is completed, we run the Bradley-Terry [1] aggregation from the Crowd-Kit [21] library for Python to obtain a ranked list of keyword sets for each image description. The final evaluation metric used in the genetic algorithm to produce the new candidate sets is the average rank of a keyword set (as described in Section 2). We use 60 image descriptions for optimization (ten from each category) and 12 for the validation of the optimization results.

For the keywords optimization, we use a genetic algorithm as follows. We parameterized every keyword set by a binary mask of length 100, indicating whether the keyword should be appended to the prompt. We initialized the algorithm with all zeros and the mask including the 15 most popular keywords. At the selection step, we took the two masks with the highest average rank. At the crossover step, we swapped a random segment of them. At the mutation step, we swapped bits of the resulting offsprings with probability of 1% to get the resulting candidates.

4.2 Results

We ran the optimization for 56 iterations on 60 image descriptions since we have a fixed annotation budget. To ensure that our method did not overfit, we ran the evaluation on another 12 descriptions (validation). Figure 4 shows ranks of tried keyword sets. According to the evaluation results in Table 1, we found that our algorithm was able to find a significantly better set of keywords than the fifteen most popular ones (Top-15). Also, we see that any set of prompt keywords is significantly better than no keywords at all (No Keywords).

[3] https://toloka.ai/

Table 1: Average rank of the baseline keywords (top-15 most common on Stable Diffusion Discord) and the ones found by the genetic algorithm. Rank is averaged over 60 prompts on train and over 12 prompts on validation (val); maximal rank is 56.

Train				Validation			
No Keywords	Top-15	Best Train	Best Val	No Keywords	Top-15	Best Train	Best Val
3.5	14.25	**43.60**	39.32	5.42	12.50	38.00	**46.00**

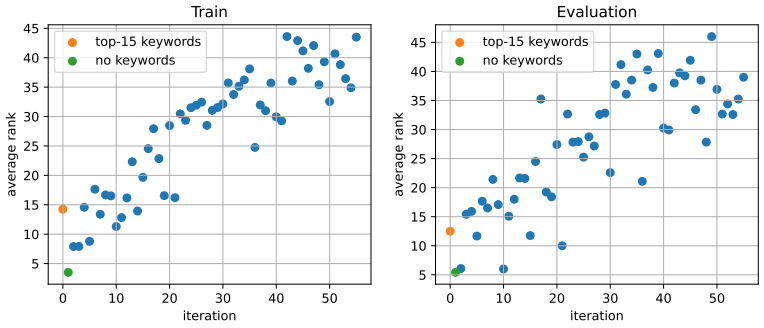

Figure 4: Average ranks of keyword sets tried by the genetic algorithm. There are total 56 keyword sets, so the maximal average rank is 56.

We see that most results hold on the validation set, too, but the metrics have more noise. Overall, the best set of keywords on the training set of 60 prompts is *cinematic, colorful background, concept art, dramatic lighting, high detail, highly detailed, hyper realistic, intricate, intricate sharp details, octane render, smooth, studio lighting, trending on artstation.* An example of images generated with this keyword set is shown in Figure 1.

4.3 Discussion

We show that adding the prompt keywords significantly improves the quality of generated images. We also noticed that the most popular keywords do not result in the best-looking images. To estimate the importance of different keywords, we trained a random forest regressor [2] on the sets of keywords and their metrics that is similar to W&B Sweeps.[4] We found that the most important keywords, in reality, are different from the most widely used ones, such as "trending on artstation." The most important keyword we found was "colorful background."

There are several limitations to our approach. We can not conclude that the found set of keywords is the best one since the genetic algorithm can easily fall into a local minimum. In our run, it tried only 56 keywords out of the 100 most popular ones. Also, our evaluation metrics are based on ranks, not absolute scores, so they are not sensitive enough to determine the convergence of the algorithm.

However, since we release all the comparisons, generated images, and code, it is possible for the community to improve on our results. For instance, one can run a genetic algorithm from a different initialization, for a larger number of iterations, or even with more sophisticated optimization methods. This can easily be done by comparing the new candidates with our images and adding these results to the dataset.

[4]https://docs.wandb.ai/guides/sweeps

5 RELATED WORK

The aesthetic quality evaluation is one of the developing topics in computer vision. There are several datasets and machine learning methods aiming at solving this problem [18, 22]. However, the available datasets contain human judgments on image aesthetics scaled from 1 to 5. Our experience shows that the pairwise comparisons that we used in this paper are a more robust approach as different humans perceive scales differently and subjectively. Also, they specify training a model to evaluate the aesthetics but not on the generative models. Large language models, such as GPT-3 [3], have enabled a wide range of research tasks on prompt engineering [5, 6, 8, 10, 12, 15, 17]. Recent papers also discover the possibilities of prompt engineering for text-to-image models and confirm that prompts benefit from the added keywords [7]. To the best of our knowledge, we are the first to apply it to find the best keywords.

6 CONCLUSION

We presented an approach for evaluating the aesthetic quality of images produced by text-to-image models with different prompt keywords. We applied this method to find the best keyword set for Stable Diffusion and showed that these keywords produce better results than the most popular keywords used by the community. Despite the fact that our work focuses on the evaluation of keywords for text-to-image models, it is not limited by this problem and can be applied for an arbitrary prompt template evaluation, for example, in the text-to-text setting. This is a direction for our future work. Last but not least, we would like to encourage the community to continue our experiment and find better keyword sets using our open-source code and data.[5]

[5]https://github.com/toloka/BestPrompts

REFERENCES

[1] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. https://doi.org/10.2307/2334029

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Tom Brown et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Curran Associates, Inc., Montréal, QC, Canada, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. 2021. DALL-E Mini. https://doi.org/10.5281/zenodo.6682042

[5] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP 2021)*. Association for Computational Linguistics, Online, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

[6] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9 (2022), 35 pages. https://doi.org/10.1145/3560815

[7] 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 (CHI '22)*. Association for Computing Machinery, New Orleans, LA, USA, 23 pages. https://doi.org/10.1145/3491102.3501825

[8] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. arXiv:2104.08786.

[9] Lucas Maystre and Matthias Grossglauser. 2017. Just Sort It! A Simple and Effective Approach to Active Preference Learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017, Vol. 70)*. PMLR, Sydney, NSW, Australia, 2344–2353. https://proceedings.mlr.press/v70/maystre17a.html

[10] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2022)*. Association for Computational Linguistics, Dublin, Ireland, 3470–3487. https://doi.org/10.18653/v1/2022.acl-long.244

[11] Nikita Pavlichenko and Dmitry Ustalov. 2021. IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons. , 5 pages. arXiv:2110.14990.

[12] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How Context Affects Language Models' Factual Predictions. arXiv:2005.04611.

[13] 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 (ICML 2021, Vol. 139)*. PMLR, Virtual Only, 8748–8763.

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

[15] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, Yokohama, Japan, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 10684–10695. https://doi.org/10.1109/cvpr52688.2022.01042

[17] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. , 2655–2671 pages. https://doi.org/10.18653/v1/2022.naacl-main.191

[18] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. 2018. Attention-Based Multi-Patch Aggregation for Image Aesthetic Assessment. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. Association for Computing Machinery, Seoul, Republic of Korea, 879–886.

[19] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015, Vol. 37)*. PMLR, Lille, France, 2256–2265.

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. arXiv:2010.02502.

[21] Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2023. Learning from Crowds with Crowd-Kit. arXiv:2109.08584 [cs.HC] https://arxiv.org/abs/2109.08584

[22] Bo Zhang, Li Niu, and Liqing Zhang. 2021. Image Composition Assessment with Saliency-augmented Multi-pattern Pooling. arXiv:2104.03133.