Learning Prompt-Level Quality Variance for Cost-Effective Text-to-Image Generation

Dongkeun Lee
Korea University
Seoul, Republic of Korea
dklee98@korea.ac.kr

Wonjun Lee
Korea University
Seoul, Republic of Korea
wlee@korea.ac.kr

Abstract

Text-to-image generation is a multivariable process in which the resulting quality is determined by both the generative model and the input prompt. While previous efforts rely on a single model either by enhancing its capability or by reformulating prompts, we point out that no single model excels at handling all types of tasks, as there exist *inter-model* and *intra-model* quality variance induced by the difference in types of prompts. This paper explores the relationship between the generation quality of text-to-image models and the linguistic features of input prompts by measuring the performance of state-of-the-art models using five different prompt datasets each with its distinctive features. Motivated by our empirical observations, we propose a novel approach that assigns each prompt to its best-performing model based on quality prediction. This enables utilizing a diverse set of models each with its expertise and cost, thereby enhancing cost-effectiveness. Evaluation results show that our approach can reduce the total generation cost by 29.25% with comparable or even higher generation quality than using only the single best model.

CCS Concepts

• **Information systems** → **Multimedia information systems**.

Keywords

generative models, text-to-image generation, cost-effectiveness

ACM Reference Format:
Dongkeun Lee and Wonjun Lee. 2024. Learning Prompt-Level Quality Variance for Cost-Effective Text-to-Image Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679954

1 Introduction

Recent studies on text-to-image generation and the unprecedented capabilities of state-of-the-art models have enabled users to visualize their wants based on prompts, which are instructions given as textual descriptions of the target image. With myriads of text-to-image models available today, diverse factors affect their resulting quality, such as the model capability and the linguistic features of the prompt. We point out that **no single model** excels at all types of tasks due to *inter-model* and *intra-model* quality variance induced by the difference in types of prompts, envisioning opportunities for enhancing cost-effectiveness that are yet unexplored.

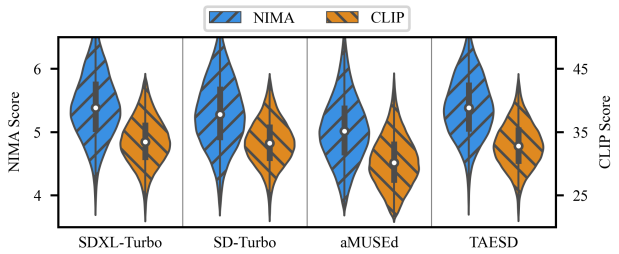

Figure 1: Quality measurement of images generated from diverse text prompts. Both aesthetic quality (NIMA Score) and text-image alignment (CLIP Score) show notable variance along with the difference between model performance.

The main reasons for *inter-model* quality variance are twofold: model properties and training data. Many text-to-image models use pre-trained language models (e.g., T5 [21] or CLIP text encoder [20]) to extract text embeddings from input prompts and are composed of diverse generative models such as generative adversarial networks (GAN) [15, 25, 30], diffusion models [4, 22, 23], autoregressive models [5, 33], or masked image models [2, 17] that process those embeddings in latent space for conditional generation. In addition to these differences between pipeline components, types and distribution of training data may vary, leading to different domain generalization and zero-shot transfer abilities. It is worth noting that the model parameter size [8, 33] also affects these capabilities. Overall, the model performance depends on how it is shaped and trained, resulting in varying generation quality across text-to-image models even with identical generation requests.

Alongside the quality variance among models, a single model may also show variance in generation quality, i.e., *intra-model* quality variance, depending on the type of task. As shown in Fig. 1, all models show notable variance in generation quality regardless of their average performance. While SDXL-Turbo shows the best performance "on average", this does not necessarily mean it excels at handling all types of text prompts. For example, among the five evaluation benchmarks we examine, TAESD marks the highest proportion in generating images with better quality than all other models when given prompts from MS/LN-COCO, 10.38% and 4.48% more times than SDXL-Turbo, respectively (cf. Sec. 2.2).

Therefore, text-to-image generation is a multivariable process in which quality is determined by both the model that handles the generation request and the features of the request (i.e., text

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679954

Table 1: Statistics of evaluation benchmarks.

Benchmark	Number of Prompts	Number of Words / Prompt		
		Min.	Max.	Avg. ($\mu(\pm\sigma)$)
MS-COCO [11]	31,427	6	45	10.46 (±2.41)
LN-COCO [19]	8,573	6	181	40.45 (±18.75)
DrawBench [24]	200	1	51	11.68 (±9.62)
PartiPrompts [33]	1,632	1	67	9.12 (±7.34)
DiffusionDB [32]	8,168	1	217	24.31 (±16.10)

prompt) itself. Previous efforts on enhancing generation quality have focused solely on reformulating prompts [6, 12, 34], relying on a single model that is believed to be an all-rounder. Still, there are opportunities for further optimization as models with a smaller number of parameters, hence at a lower cost, may generate images with comparable or even better qualities for certain types of tasks. Utilizing these models each with its *expertise* leads to improved cost-effectiveness in the overall text-to-image generation process.

Motivated by empirical observations on prompt-level generation quality that varies both inter-model and intra-model, we propose a novel approach that assigns each prompt to its best-performing model based on quality prediction. For each generation request, we select the model that will generate an image with the highest quality based on the linguistic features of the given prompt. This enables utilizing a smaller, inexpensive model with less diversity in generation but excels at its expertise. Although there have been studies on utilizing performance variance among large language models (LLM) [3, 31], our work is the first attempt to apply this methodology in the field of text-to-image generation.

We measure the performance of state-of-the-art models using diverse evaluation benchmarks to study the feasibility of utilizing multiple models with varying capabilities and costs. We evaluate the effectiveness of our approach using different quality metrics and model selection strategies. Our results show that assigning each generation request to the most suitable model can reduce the total cost by 29.25% on average while generating images of which the overall quality is in line with the model of the highest performance. The major contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to utilize prompt-level quality variance among text-to-image models to enhance cost-effectiveness in image generation.
- We provide an empirical analysis that shows inter-model and intra-model quality variance according to the linguistic features of input prompts.
- We propose a novel approach that selects the best model for each prompt based on its linguistic features. Evaluation results show that this can reduce total generation cost by 29.25% with comparable or even higher quality outcomes.

2 Data Analysis: Input Varieties and Quality Variance

We begin by showing the results from studying quality variance induced by varieties in text prompts. Our results show the feasibility of utilizing multiple models with different capabilities and costs.

Table 2: Performance comparison between text-to-image models used in our experiment: generation quality, inference speed, and GPU memory usage.

Model (Sampling Steps)	NIMA Score ↑	CLIP Score ↑	Inf. Time ($\mu(\pm\sigma)$)	Memory Footprint
SDXL-Turbo [28] (4 steps)	5.405	33.59	0.616 s (±0.071)	9.51 GB
SD-Turbo [27] (1 step)	5.292	33.34	0.176 s (±0.018)	4.64 GB
aMUSEd [17] (12 steps)	5.024	30.09	0.489 s (±0.047)	3.75 GB
TAESD [16] (25 steps)	5.397	32.90	1.588 s (±0.053)	3.48 GB

2.1 Constructing Text-to-Image Performance Dataset

We first collect text-to-image performance dataset by generating images using state-of-the-art models with diverse sets of prompts. We evaluate the generation quality in two aspects: aesthetic quality and text-image alignment. To measure aesthetic quality, we use NIMA [10, 29], a learning-based framework that predicts whether an image is visually attractive or with good technical quality. Additionally, we calculate CLIP score [7] to measure text-image alignment using OpenCLIP ViT-g/14. To minimize the effect of randomness on image synthesis, we generate three images per prompt and compute the average quality. All models generate images of size 512×512. The machine used in this experiment has a GeForce RTX 2080 Ti GPU and an Intel i7-8700K CPU with 16 GB of RAM.

2.1.1 Evaluation Benchmarks. We build a set of evaluation benchmarks with 50k prompts from five different prompt datasets each with its distinctive features as shown in Table 1. We use the COCO validation set (MS-COCO), the standard dataset for evaluating cross-modal tasks [13, 14, 24, 33], with LN-COCO, the COCO portion of the Localized Narrative dataset that consists of longer, detailed descriptions of MS-COCO's reference images. We use all prompts in LN-COCO and randomly draw 31,427 prompts from MS-COCO.

In addition to MS/LN-COCO, of which descriptions are generally limited to common scenes and objects, we use some more challenging sets of prompts. DrawBench and PartiPrompts are designed to test different model capabilities across a range of challenging aspects such as handling complex, abstract prompts. Meanwhile, DiffusionDB consists of user-generated prompts collected by scraping prompt-image pairs. We use all prompts from DrawBench and PartiPrompts and randomly sample 8,168 prompts from DiffusionDB.

2.1.2 Text-to-Image Models. We select four state-of-the-art models by jointly considering their performance as shown in Table 2. SDXL-Turbo and SD-Turbo are each distilled version of Stable Diffusion XL (SDXL) 1.0 and Stable Diffusion v2.1 (SD2.1), trained with adversarial diffusion distillation [26] so as to preserve its original capability with reduced model size and inference time. aMUSEd is a light-weight masked image model that also features sub-second image generation. TAESD is based on SD2.1 and uses a distilled variational autoencoder. While with longer inference time, TAESD generates

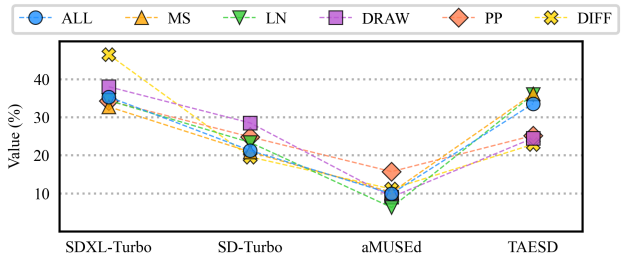

Figure 2: Proportion of each model generating images with the highest quality in terms of NIMA score when given: Total Benchmark (ALL), MS-COCO (MS), LN-COCO (LN), DrawBench (DRAW), PartiPrompts (PP), and DiffusionDB (DIFF).

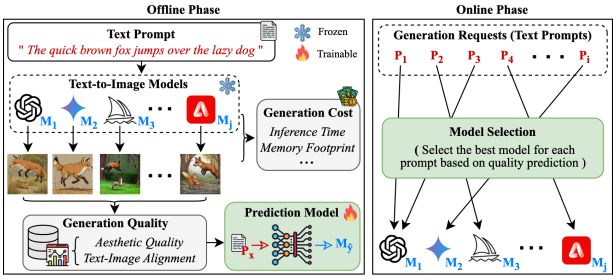

Figure 3: The architecture overview of our proposed approach for cost-effective text-to-image generation.

images of which quality is comparable to that of SDXL-Turbo and SD-Turbo, and has the smallest memory footprint, making it a viable option for cost-effective image generation.

2.2 Breakdown Analysis

We perform an in-depth analysis of the generation quality of each text-to-image model in terms of each prompt to gain a deeper insight into the quality variance induced by the difference in types of tasks. After generating images and measuring their quality with the evaluation benchmarks, we count the occurrences of each model producing images with the highest quality among other models for the given prompt. Fig. 2 shows the results stratified by datasets.

When given the total benchmark, the proportion of each model follows the results on the overall performance shown in Table 2, with SDXL-Turbo and TAESD accounting for 35.34% and 33.54% of generating best cases, respectively. SDXL-Turbo even reaches 46.46% when with prompts from DiffusionDB. However, we observe that this trend does not always hold. When given prompts from MS/LN-COCO, TAESD outperforms SDXL-Turbo, generating the highest-quality images 10.38% and 4.48% more frequently with MS/LN-COCO, respectively. We also note that there are quite a few cases where aMUSEd plays a role in generating quality images, with its proportion of generating best cases ranging from 6.28% (LN) to 11.12% (DIFF) and 15.69% (PP). This stratification over datasets, i.e., type of tasks, uncovers that there are certain tasks that each model excels at, as each set of prompts differs in its linguistic features, such as the vocabulary used or the length of prompts.

3 Learning Quality Variance for Cost-Effective Model Selection

Motivated by our observations on the variability in generation quality, we propose a novel approach that learns to select the best model based on the linguistic features of input prompts.

3.1 Framework Overview

The overall architecture of our proposed approach is depicted in Fig. 3. During the offline phase, we run performance tests using a set of evaluation benchmarks to measure the generation quality of each model for the given prompt. As shown in the samples in Fig. 3, generated images may be either low in fidelity (i.e., less aesthetically appealing) or misaligned with the user's request (e.g., absence of the lazy dog). Therefore, as discussed in Sec. 2.1, we evaluate the generation quality in terms of both aesthetic quality and text-image alignment to jointly consider these metrics in selecting the best-performing model. After dataset collection, based on the pairs of text prompts and the resulting quality measurements, we train a quality prediction model that predicts the best-performing model for the given prompt based on its linguistic features. The cost of each generation model can be defined by AI service providers, e.g., via API pricing. In this work, we set the cost of each model based on its inference time and memory footprint (cf. Sec. 4.2). During the online phase, for the series of incoming requests, our prediction model assigns each generation request to the selected text-to-image model, aiming to maximize total generation quality at a lower cost.

3.2 Quality Prediction Model

We formulate the task of prompt-level quality prediction as a classification problem of predicting which model will generate an image with the highest quality based on the given prompt.

Based on the performance dataset collected with P^B, a set of benchmark prompts, and $M = \{M_1, \ldots, M_j\}$, a set of text-to-image models each with its own capabilities, the best-performing model M_y for a text prompt P_x^B is defined as:

$$y = \underset{m \in \{1, \ldots, j\}}{\arg\max}\ Q(M_m(P_x^B)) \tag{1}$$

where Q denotes the generation quality. Using cross-entropy loss $l(\cdot)$, the quality prediction model $F(\cdot)$ is trained to minimize:

$$\sum_{P_x^B} l(F(P_x^B), M_y) \tag{2}$$

After learning the relationship between the linguistic features of input prompt and its resulting quality, for a set of generation requests $P^R = \{P_1, \ldots, P_i\}$, each request P_x^R is assigned to $M_{\hat{y}}$:

$$M_{\hat{y}} = F(P_x^R) \tag{3}$$

which is the most suitable model that is predicted to generate an image of the highest quality with P_x^R among M.

4 Evaluation Results

This section presents the results of our evaluation. Using the dataset collected as discussed in Sec. 2.1, we randomly split the entire dataset into a ratio of 8:1:1, each for the training, validation, and test set, respectively. After training the quality prediction model using

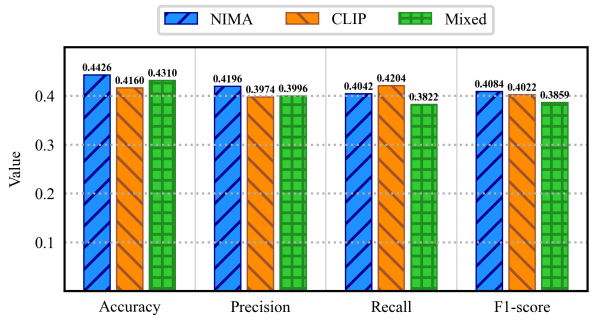

Figure 4: Prediction performance of our CLIP-based model with different quality metrics.

the training/validation set, we evaluate its prediction performance and the cost-effectiveness of our approach based on the test set.

4.1 Prediction Performance

Our quality prediction model consists of a CLIP text encoder with a classification head on top. We opt to use CLIP as it shares common features with other text encoders used in text-to-image models [17, 18, 25, 30] to extract embeddings from input prompts, thus suitable than other models such as Long Short-Term Memory (LSTM) [9] or DistilBERT [31]. We choose ViT-B/16 as the CLIP model.

In addition to NIMA score and CLIP score, we use Mixed score to jointly consider the two metrics in selecting the best-performing model for each prompt. We apply min-max normalization to each measured result so that both scores are on the same scale, then compute the sum of normalized scores. The model is trained for 10 epochs using AdamW optimizer and a learning rate of 6.4×10^{-6}.

As shown in Fig. 4, we can observe that regardless of the quality metric, our CLIP-based model can find the model that generates an image with the highest quality with more than 41% accuracy. Using NIMA score as a quality metric leads to the highest accuracy (44.26%) and precision (41.96%), while using CLIP score leads to the highest recall (42.04%). Although using Mixed score shows similar results overall, it results in lower performance than using a single metric as selection criteria, with the lowest F1-score at 38.59%. We will discuss more about this result in the following Sec. 4.2.

We also note that finding the best-performing model is not an all-or-nothing affair. A sub-optimal selection may still be of value in cases generating images with comparable quality or at a lower cost. Indeed, when using Mixed score, we observe that 51.53% of those sub-optimal selections generate images with the second-highest quality, taking part in the overall quality enhancement.

4.2 Cost Effectiveness

We compare our approach with the following strategies by evaluating the average generation quality and the total cost.

- **Oracle**: The optimal assignment on the ground-truth data by always selecting the model with the highest quality outcome.
- **Single-model**: Assigning a single, fixed model for every generation request. We evaluate this strategy with each available model in our experiment.
- **CEMS**: Our approach for Cost-Effective Model Selection.

Table 3: The results of average quality and total cost obtained when applying each model selection strategy on the test set with NIMA, CLIP, and Mixed score as selection criteria. † refers to our proposed approach.

Strategy	NIMA Score		CLIP Score		Mixed Score		
	NIMA ↑	Cost ↓	CLIP ↑	Cost ↓	NIMA ↑	CLIP ↑	Cost ↓
Oracle	5.625	0.3876	35.16	0.3461	5.562	34.47	0.3864
SDXL-Turbo	5.405	0.5133	33.66	0.5133	5.405	33.66	0.5133
SD-Turbo	5.303	0.0733	33.40	0.0733	5.303	33.40	0.0733
aMUSEd	5.034	0.1630	30.13	0.1630	5.034	30.13	0.1630
TAESD	5.401	0.5293	32.92	0.5293	5.401	32.92	0.5293
CEMS †	5.462	0.3833	33.75	0.3476	5.434	33.60	0.3586

Motivated by the pricing model of serverless computing [1], we set the cost of each generation request as follows:

$$\text{Inference Time (s)} \times \lceil \text{Memory Footprint (GB)} \rceil \times 0.0000166667 \quad (4)$$

For the inference time of each model, we use the mean value as it shows a minor variance of less than 0.1 s (Table 2).

Table 3 shows the details of the results. When compared with SDXL-Turbo, the model with the highest average performance, our approach reduces the total cost by 29.25% on average while achieving higher generation quality in almost all cases, with its largest reduction of 32.28% using CLIP score as selection criteria. Even when compared with the Oracle, albeit lower in average quality than the optimal assignment, our approach costs less when using NIMA score or Mixed score as selection criteria, demonstrating its cost-effectiveness. It is noteworthy that this is by virtue of utilizing inexpensive models: SD-Turbo or aMUSEd, that can generate images of similar caliber but at a considerably lower cost.

Similar to the results in Fig. 4, using Mixed score results in a minor drop in average quality both for our approach and the Oracle. We attribute this result to the non-linear relationship between NIMA score and CLIP score, with a Pearson correlation of 0.1883. Addressing this challenge will be explored in our future work.

5 Conclusion

In this paper, we studied the relationship between the generation quality of text-to-image models and the linguistic features of input prompts. We showed diverse types of quality variance induced by varieties in text prompts and proposed a novel approach that selects the best-performing model for each generation request based on quality prediction. The experimental results demonstrated the cost-effectiveness of our approach, reducing total generation cost by up to 32.28% while resulting in quality outcomes that are on par with utilizing a single model of the highest average performance.

Acknowledgments

This work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00234719, SW Star Lab) for the Service Continuity-Oriented Edge Continuum SW Framework and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00338786).

References

[1] Amazon Web Services. [n. d.]. AWS Lambda Pricing. Retrieved May 24, 2024 from https://aws.amazon.com/lambda/pricing/

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv:2301.00704 [cs.CV]* (2023).

[3] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv:2305.05176 [cs.LG]* (2023).

[4] Zhongjie Duan, Chengyu Wang, Cen Chen, Jun Huang, and Weining Qian. 2023. Optimal Linear Subspace Search: Learning to Construct Fast and High-Quality Schedulers for Diffusion Models. In *Proc. of ACM CIKM*. Birmingham, United Kingdom, 463–472.

[5] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Proc. of ECCV*. Tel Aviv, Israel, 89–106.

[6] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing Prompts for Text-to-Image Generation. *arXiv:2212.09611 [cs.CL]* (2022).

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proc. of EMNLP*. Online and Punta Cana, Dominican Republic, 7514–7528.

[8] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up GANs for Text-to-Image Synthesis. In *Proc. of IEEE/CVF CVPR*. Vancouver, BC, Canada, 10124–10134.

[9] Dongkeun Lee, Minwoo Joo, and Wonjun Lee. 2023. Net-track: Generic Web Tracking Detection Using Packet Metadata. In *Proc. of ACM WWW*. Austin, TX, USA, 2230–2240.

[10] Christopher Lennan, Hao Nguyen, and Dat Tran. 2018. Image Quality Assessment. Retrieved May 18, 2024 from https://github.com/idealo/image-quality-assessment

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of ECCV*. Zurich, Switzerland, 740–755.

[12] Yinqiu Liu, Hongyang Du, Dusit Niyato, Jiawen Kang, Shuguang Cui, Xuemin Shen, and Ping Zhang. 2023. Optimizing Mobile-Edge AI-Generated Everything (AIGX) Services by Prompt Engineering: Fundamental, Framework, and Case Study. *arXiv:2309.01065 [cs.NI]* (2023).

[13] Yun Liu, Xiaoming Zhang, Feiran Huang, and Zhoujun Li. 2018. Adversarial Learning of Answer-Related Representation for Visual Question Answering. In *Proc. of ACM CIKM*. Torino, Italy, 1013–1022.

[14] Junyu Luo, Ying Shen, Xiang Ao, Zhou Zhao, and Min Yang. 2019. Cross-modal Image-Text Retrieval with Multitask Learning. In *Proc. of ACM CIKM*. Beijing, China, 2309–2312.

[15] Junyeong Maeng, Kwanseok Oh, and Heung-Il Suk. 2023. Age-Aware Guidance via Masking-Based Attention in Face Aging. In *Proc. of ACM CIKM*. Birmingham, United Kingdom, 4165–4169.

[16] Ollin Boer Bohan. 2023. Tiny AutoEncoder for Stable Diffusion. Retrieved May 22, 2024 from https://github.com/madebyollin/taesd

[17] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. 2024. aMUSEd: An Open MUSE Reproduction. *arXiv:2401.01808 [cs.CV]* (2024).

[18] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proc. of ICLR*. Vienna, Austria, 1–13.

[19] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting Vision and Language with Localized Narratives. In *Proc. of ECCV*. Virtual Event, 1–24.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs.CV]* (2021).

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[22] 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]* (2022).

[23] 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]* (2022).

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv:2205.11487 [cs.CV]* (2022).

[25] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv:2301.09515 [cs.LG]* (2023).

[26] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial Diffusion Distillation. *arXiv:2311.17042 [cs.CV]* (2023).

[27] Stability AI. 2023. SD-Turbo Model Card. Retrieved May 20, 2024 from https://huggingface.co/stabilityai/sd-turbo

[28] Stability AI. 2023. SDXL-Turbo Model Card. Retrieved May 20, 2024 from https://huggingface.co/stabilityai/sdxl-turbo

[29] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.* 27, 8 (2018), 3998–4011.

[30] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In *Proc. of IEEE/CVF CVPR*. Vancouver, BC, Canada, 14214–14223.

[31] Marija Šakota, Maxime Peyrard, and Robert West. 2024. Fly-Swat or Cannon? Cost-Effective Language Model Choice via Meta-Modeling. In *Proc. of ACM WSDM*. Merida, Mexico, 606–615.

[32] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A Large-Scale Prompt Gallery Dataset for Text-to-Image Generative Models. In *Proc. of ACL*. Toronto, ON, Canada, 893–911.

[33] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *arXiv:2206.10789 [cs.CV]* (2022).

[34] Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jia Chen, and Shaoping Ma. 2024. Capability-aware Prompt Reformulation Learning for Text-to-Image Generation. In *Proc. of ACM SIGIR*. Washington DC, USA, 2145–2155.