A Additional Details of SelfEval

In this section, we provide a detailed algorithm and systematic figure of SelfEval in Algorithm 1 and Figure 1 respectively. SelfEval iteratively denoises an image, similar to the reverse process of diffusion models, but instead estimates the likelihood of an image-text pair.

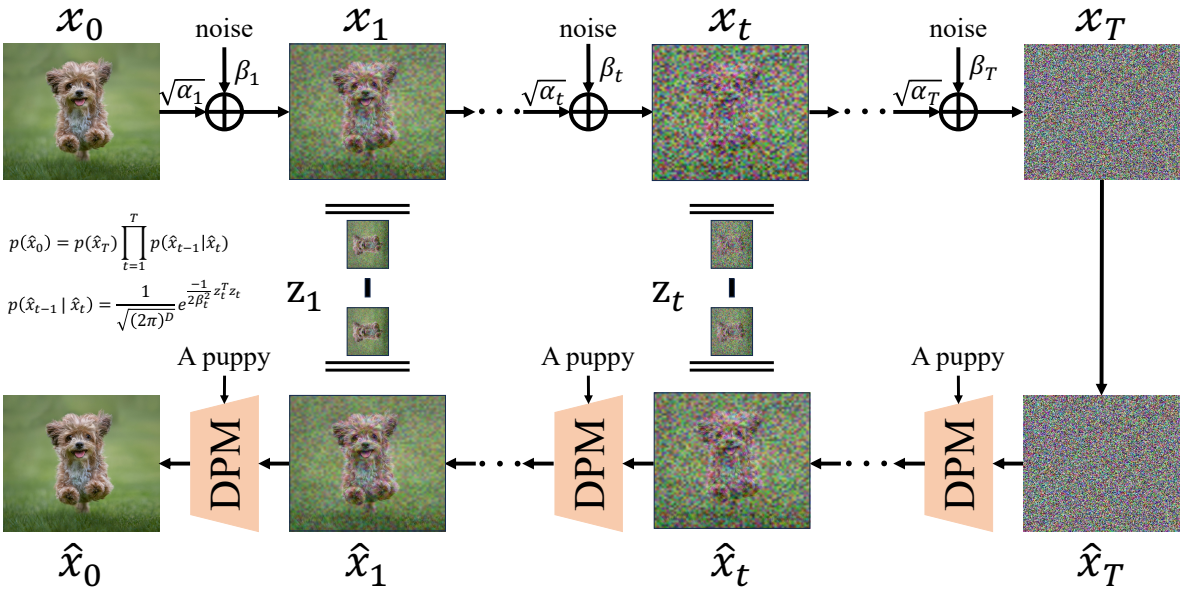

Figure 1: Illustration of proposed method: (Left) Starting from a noised input, the standard diffusion sampling method denoises the input iteratively to generate images from the input distribution. (Middle): SelfEval takes an image x_0 and conditioning c pairs to estimates the likelihood $p(x_0|c)$ of the pair in an iterative fashion. (Right): Given an image, x_0 and n captions, $\{c_0, c_1, \ldots, c_n\}$, SelfEval is a principled way to convert generative models into discriminative models. In this work, we show that the classification performance of these classifiers can be used to evaluate the generative capabilities.

Algorithm 1 Algorithm for estimating $p(\mathbf{x}|\mathbf{c})$ using SelfEval

Input: Diffusion model $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$; Input image $\mathbf{x}_0$; Forward latents: $\{\mathbf{x}_{1:T}\}$; Reverse latents: $\{\hat{x}_{1:T}\}$; Number of trials: N
for i=1:N **do**
  Sample noise $\sim \mathcal{N}(0, \mathbb{I})$
  $\mathbf{x}_{1:T} = q_{\text{sample}}(\mathbf{x}_0, t = 1:T, \text{noise} = \text{noise})$; $\mathbf{x}_t \in \mathbb{R}^D$
  conditionals $\longleftarrow$ [ ]
  for j=1:T **do**
    $p(\mathbf{x}_{t-1}|\bar{\mathbf{x}}_t, \mathbf{c}) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_\theta|}} e^{-0.5(\mathbf{x}_{t-1}-\boldsymbol{\mu}_\theta(\bar{\mathbf{x}}_t,t,\mathbf{c}))^T \mathbf{\Sigma}_\theta^{-1}(\mathbf{x}_{t-1}-\boldsymbol{\mu}_\theta(\bar{\mathbf{x}}_t,t,\mathbf{c}))}$
    conditionals = [conditionals ; $p(\mathbf{x}_{t-1}|\bar{\mathbf{x}}_t, \mathbf{c})$]
  end for
  Compute $p(\mathbf{x}_T) = \frac{1}{\sqrt{(2\pi)^D}} e^{\frac{-1}{2\beta_T^2}\|\mathbf{x}_T\|^2}$
  Compute likelihood $p_i(\mathbf{x}_0|\mathbf{c}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p(\mathbf{x}_{t-1}|\bar{\mathbf{x}}_t, \mathbf{c})$
end for
$p(\mathbf{c}|\mathbf{x}_0) = \frac{p(\mathbf{x}_0|\mathbf{c})}{|\mathcal{C}|}$

B Details of Human evaluation

Human evaluations are the de-facto standard for judging the performance of text-to-image models. we adopt a conventional A/B testing approach, wherein raters are presented with generations from two models and are asked to vote for one of four choices: "both" the generations are faithful, "none" of them are faithful, or if only one of the two models ("model 1" or "model 2") demonstrates fidelity to the given prompt. We show the template provided to the raters in Figure 2. The template includes three examples that advice the raters on how to rate a given sample followed by a text prompt and two images. The four possible choices are shown on the right in Figure 3. The images used as instructions for the human raters are shown in Figure 3. Figure 3 shows three pairs of images with the text prompt below them. The first example shows two images that are faithful to the input prompt but the quality of one (left) image superior to the other (right). Since, we ask the raters to evaluate the text faithfulness, we recommend picking the "both" option for such samples. The second image shows an example where only one of the images is faithful to the text. The raters are instructed to pick the option corresponding to the right image in this case. The final example shows two images that are not faithful to the text prompt. The raters are adviced to pick the "none" option in this scenario.

C Ablation Experiments

Table 1: Effect of timesteps on the performance of SelfEval on the six splits.

T	Attribute	Color	Count	Shape	Spatial	Text Corruption
50	54.2	32.2	26.3	34.9	33.0	25
100	54.3	34	25.8	30.2	38.0	24.3
250	53	32.3	27.4	35	32.7	21.7

Table 2: Effect of N on the performance of SelfEval on the six splits.

N	Attribute	Color	Count	Shape	Spatial	Text Corruption
1	53.0	26.0	27.2	35.2	31.2	20.7
5	54.3	31.7	25.7	34.9	33.0	22.1
10	54.3	34.0	25.8	32.5	38.6	24.3
15	53.4	36.3	28.0	36.3	32.8	22.8

Table 3: Effect of the choice of seed on the performance of SelfEval.

S	Attribute	Color	Count	Shape	Spatial	Text Corruption
1	54.3	34.0	25.8	32.5	38.6	24.3
2	53.0	26.0	27.2	35.2	31.2	20.7
3	54.3	31.70	25.7	34.9	33.0	22.1
std	0.5	0.5	0.9	1.4	1.5	0.8

In this section we analyze the effect of various components that affect the performance of SelfEval on the six splits introduced in the main paper. We use the LDM-T5 model for all our experiments.

Effect of T: SelfEval has a time complexity of $\mathcal{O}(NT)$ and Table 1 shows the the effect of timesteps on the performance of SelfEval. We observe that SelfEval achieves the best result at different timesteps for different datasets. We notice that the performance drops as we increase the timesteps from 100 to 250 in most cases. As the number of timesteps increases, we believe that the fraction of them responsible for text faithfulness decrease, resulting in a drop in performance. We find $T = 100$ to be a good tradeoff for performance and speed and is used for all the experiments on the six data splits.

Effect of N: Table 2 shows the results of the effect of number of trials N on the performance of SelfEval. We observe that $N = 10$ works best across all the six splits and is the default choice for N.

Effect of seeds: SelfEval corrupts an input image using standard gaussian noise in each trial and we analyze the effect of the seed on the performance of SelfEval in Table 3. We observe that the performance is stable across all the six splits with a standard deviation within 1 percentage point in most of the cases. We report the seed number instead of the actual value for brevity and use the seed 1 as the default choice for all the experiments.

D Additional experiments on Winoground

In this section we ablate a few design decisions on the Winoground dataset. We use the LDM-T5 model for all the experiments.

Effect of T: We show the effect of the number of timesteps on the performance of SelfEval on the Winoground dataset in Table 4. From Table 4, we observe that SelfEval achieves the best

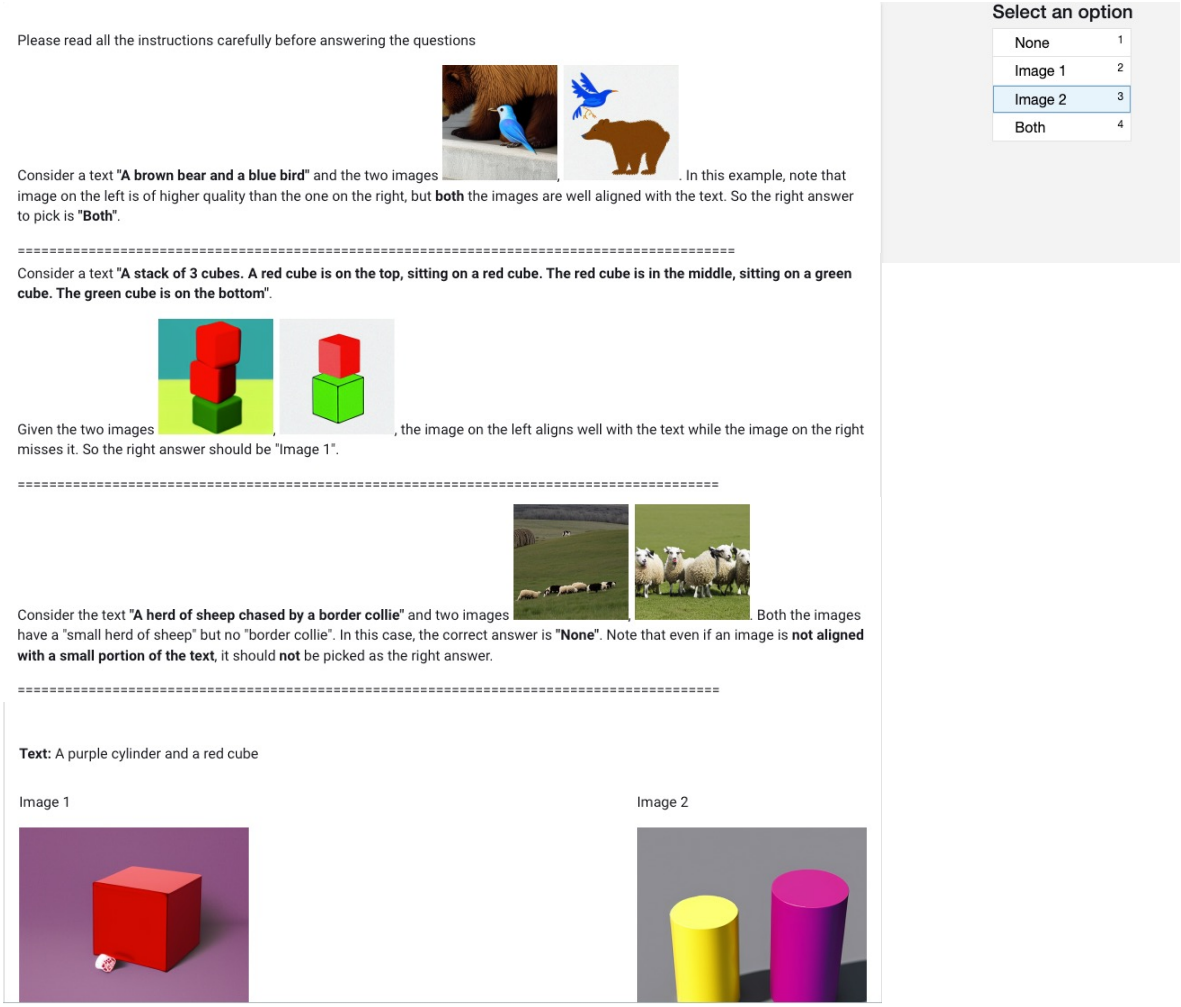

Figure 2: Template for Human raters. The template consists of instructions explaining the nature of the task (top) followed by a text prompt with two generations (bottom). Humans are expected to pick one of four options (shown on the right): "both" the generations are faithful, "none" of them are faithful, or if only one of the two images ("Image 1" or "Image 2") demonstrates fidelity to the text prompt.

Table 4: Effect of timesteps on the performance of SelfEval on the Winoground dataset

T	Image Score	Text Score
20	11.50	30.75
50	13.50	29.00
100	12.25	25.25
250	11.25	27.75

Table 5: Effect of the # of trials on the performance of SelfEval on the Winoground dataset

N	Image Score	Text Score
1	17.00	26.25
5	14.75	26.00
10	13.50	29.00
20	11.25	24.75

Table 6: Effect of the choice of seed on the performance of SelfEval on the Winoground dataset

S	Image Score	Text Score
1	13.50	29.00
2	13.00	27.00
3	12.00	28.50
	12.83± 0.76	28.17±1.04

result for image and text score at different time steps. Image score is a harder task compared to Text score Thrush et al. (2022) and hence SelfEval needs more timesteps to perform better on Image score. As the number of timesteps increase, we observe a drop in both Image and Text scores. Studies Li et al. (2023b) show that the earlier timesteps generate low frequency information (responsible for text

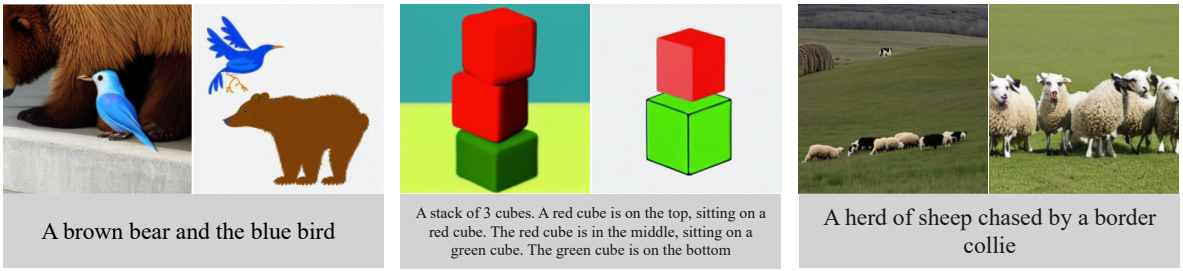

Figure 3: Instructions for Human raters. We provide three examples describing all the possible scenarios. The first example shows two images that are faithful to the text but with varying image qualities. To prevent the raters from conflating image quality with text faithfulness, we recommend the raters to pick "both" for such examples. The second example illustrates a case where only one of the image is faithful to the text. In this case, the raters are adviced to pick the option corresponding to the right image ("Image 1" in this case). The final example shows a case where both the examples are not faithful to the text (there is no border collie), in which case, we advice the raters to pick "none".

fidelity), while latter ones are responsible for high frequency appearance details. By increasing the number of timesteps, the fraction of timesteps contributing to improving the faithfulness to text (and thereby image and text scores) decreases, resulting in a drop in performance. All other experiments on Winoground use T=50 unless otherwise specified.

Effect of N: We show the effect of the number of trials (N) in Table 5. With fewer trials, the estimates are not reliable and larger trials make it computationally expensive. We observe that we attain a good tradeoff for performance and speed with $N = 10$.

Effect of the seed: We show the effect of seed on the performance of SelfEval in Table 6. We just report the seed number for brevity. We observe that both the scores are relatively stable across different values of seed. We fix seed #1 for all the experiments in this work.

E Converting COCO image-caption pairs for ITM

We use image-caption pairs from COCO for the tasks of `Color`, `Count` and `Spatial relationships`. We use the question answering data collected by authors of TIFA Hu et al. (2023) to construct data for our tasks. We pick only samples constructed from COCO. Given question answering samples from TIFA, we identify the corresponding image-caption pair from COCO and replace the correct answer in the caption with the multiple choices to form samples for the task of Image-Text Matching.

F Limitations

SelfEval relies on the sampling of the generative model to compute the scores. So the limitations of the sampling process of a generative model affect SelfEval. To be precise, for a model with T diffusion time steps and a classification task with C classes, SelfEval samples N noise signals. This results in an overall complexity of the order $\mathcal{O}(NCT)$ for computing probabilities using SelfEval. The complexity increases linearly with the number of classes C making it difficult to scale to thousands of classes (like ImageNet Deng et al. (2009)). However, several optimizations like randomly picking a starting timestep to denoise (instead of all T diffusion timesteps) and efficient classification tricks Li et al. (2023a) can be employed to improve the time complexity of SelfEval. Additionally, unlike other black-box evaluation methods, which only require the generations from the model, SelfEval requires the model definition and its checkpoints for evaluation making it impossible to evaluate closed-source generative models without model definition and checkpoint.

References

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *NeurIPS*, 2023. 9, 11, 12, 13

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023. URL https://openreview.net/forum?id=laWYA-LXlNb. 4

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 21

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. URL https://api.semanticscholar.org/CorpusID:234357997. 3

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139 – 144, 2014. URL https://api.semanticscholar.org/CorpusID:12209503. 2

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2022. URL https://api.semanticscholar.org/CorpusID:253734854. 11

Xuehai He, Weixi Feng, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Discriminative diffusion models as few-shot vision and language learners, 2023. 4

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595. 9

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. 2, 3

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 7, 21

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below. 8, 12

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2016. URL https://api.semanticscholar.org/CorpusID:15458100. 7

Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual Information Divergence: A unified metric for multimodal generative models. In *Advances in Neural Information Processing Systems 35*, 2022. URL https://openreview.net/forum?id=wKd2XtSRsjl. 2, 9, 11, 12

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114. 2

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, may 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7. 7

Martha Lewis, Nihal V. Nayak, Qinan Yu, Jack Merullo, and Elizabeth-Jane Pavlick. Does clip bind concepts? probing compositionality in large image models. *ArXiv*, abs/2212.10537, 2022. URL https://api.semanticscholar.org/CorpusID:254877746. 7

Alexander Cong Li, Mihir Prabhudesai, Shivam Duggal, Ellis Langham Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023a. URL https://openreview.net/forum?id=Ck3yXRdQXD. 4, 9, 10, 11, 21

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, 2023b. URL https://api.semanticscholar.org/CorpusID:255942528. 3, 20

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. 7, 12

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *ICCV*, 2023. 3

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation, 2023. 9, 11, 12, 13

Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification, 2023. 4

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html. 3

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=bKBhQhPeKaF. 4, 12

William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023. 3

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. URL https://api.semanticscholar.org/CorpusID:11758569. 2

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html. 3, 8

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3, 8

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 3, 8, 11

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. URL https://api.semanticscholar.org/CorpusID:248986576. 3, 7, 8, 11

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html. 3, 4

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP. 8

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159. 9

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023. 11

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022. URL https://api.semanticscholar.org/CorpusID:248006414. 4, 10, 12, 20

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *ArXiv*, abs/2401.11817, 2024. URL https://api.semanticscholar.org/CorpusID:267069207. 2

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX. 7, 9