Supplementary Material

1. Details of Reviewed Papers

We count the number of papers in which a certain experimental detail is reported. Tab. 1 shows that many papers fail to describe important details including the number of annotations and the number of ratings per sample. For annotation quality assessment, no paper report inter-annotator-agreement. The number of papers that employ certain types of evaluation criteria and rating method are also summarized in Tab. 1. We find that evaluation criteria and how to collect ratings vary from one paper to another. The full list of surveyed papers is in Tab. 2.

Table 1. The number of papers that report the details. Critical details are often omitted. The way of rating varies by paper.

		Count
Numbers	# samples	18
	# raters	11
	# rates / sample	4
Task Design	Question	20
	Label	20
	Instruction	5
Quality check	IAA	0
Annotator pool	Crowdsourcing	3
	NA	17
Crowdsourcing parameters	qualifications	2
	compensations	3
Criteria	Quality	18
	Relevance to prompts	14
	Others	2
Types of rating	Choice (w/wo ties)	10
	Ranking	9
	Numeric	3

Table 2. Full list of surveyed papers.

Title	Year	Venue
An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion [6]	2022	ArXiv
AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks [27]	2018	CVPR
CogView: Mastering Text-to-Image Generation via Transformers [2]	2021	NeurIPS
CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers [3]	2022	ArXiv
Controllable Text-to-Image Generation [11]	2019	NeurIPS
CookGAN: Causality Based Text-to-Image Synthesis [36]	2020	CVPR
CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis [14]	2020	ECCV
Cross-Modal Contrastive Learning for Text-to-Image Generation [31]	2021	CVPR
DAE-GAN: Dynamic Aspect-Aware GAN for Text-to-Image Synthesis [21]	2021	ICCV
DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis [25]	2022	CVPR
DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis [37]	2019	CVPR
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation [22]	2022	ArXiv
Dual Adversarial Inference for Text-to-Image Synthesis [10]	2019	ICCV
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models [17]	2022	ArXiv
High-Resolution Image Synthesis With Latent Diffusion Models [20]	2022	CVPR
Imagic: Text-Based Real Image Editing with Diffusion Models [9]	2022	ArXiv
Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis [8]	2018	CVPR
Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors [5]	2022	ECCV
MirrorGAN: Learning Text-To-Image Generation by Redescription [18]	2019	CVPR
Object-Driven Text-To-Image Synthesis via Adversarial Training [12]	2019	CVPR
Photographic Text-to-Image Synthesis With a Hierarchically-Nested Adversarial Network [34]	2018	CVPR
Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding [23]	2022	ArXiv
RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis From Prior Knowledge [1]	2020	CVPR
Scaling Autoregressive Models for Content-Rich Text-to-Image Generation [30]	2022	ArXiv
Semantics Disentangling for Text-To-Image Generation [29]	2019	CVPR
Semantics-Enhanced Adversarial Nets for Text-to-Image Synthesis [24]	2019	ICCV
StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks [32]	2017	ICCV
StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks [33]	2018	ArXiv
StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation [16]	2022	ECCV
StyleT2I: Toward Compositional and High-Fidelity Text-to-Image Synthesis [13]	2022	CVPR
Text to Image Generation With Semantic-Spatial Aware GAN [15]	2022	CVPR
Text-to-Image Synthesis Based on Object-Guided Joint-Decoding Transformer [26]	2022	CVPR
TISE: Bag of Metrics for Text-to-Image Synthesis Evaluation [4]	2022	ECCV
Towards Language-Free Training for Text-to-Image Generation [35]	2022	CVPR
Trace Controlled Text to Image Generation [28]	2022	ECCV
Vector Quantized Diffusion Model for Text-to-Image Synthesis [7]	2022	CVPR
Zero-Shot Text-to-Image Generation [19]	2021	PMLR

2. Annotation interface

We show screenshot of our instructions for the annotation task (Fig. 2), annotation interface (Fig. 1), and pre-task qualification test for *skillfulness* qualification (Fig. 3). The implementation of the interfaces will be published.

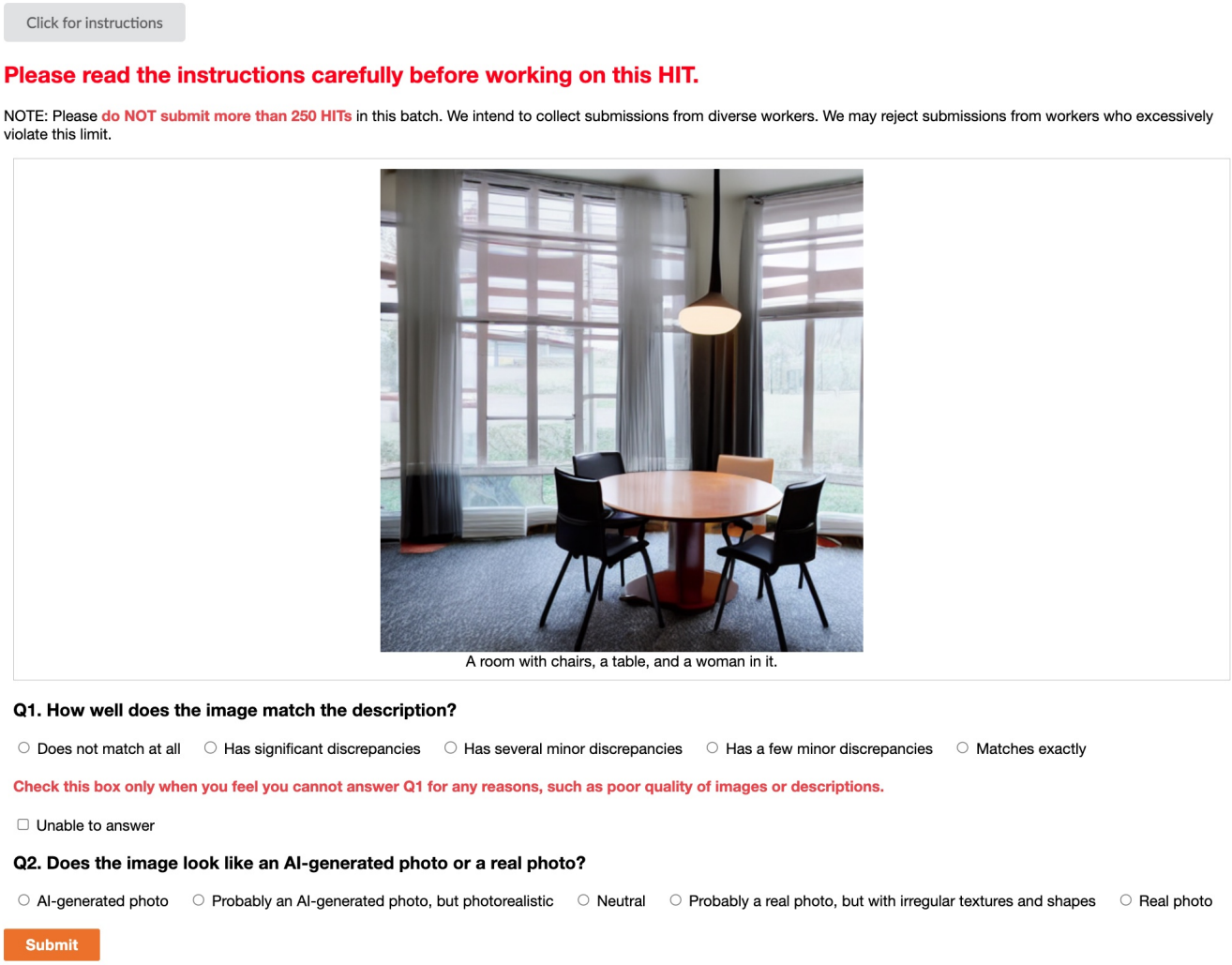


Figure 1. Annotation interface. Annotators rate the image in terms of fidelity and alignment.

Instructions

Summary | Detailed Instructions | Examples

Instruction

1. Read the description below the image carefully.
2. Check the content of the AI-generated image.
3. (Q1) Rate if the image fully represents the content of the caption.
4. (Q2) Check if any irregular shapes and textures are found in the image and rate the quality of the image.

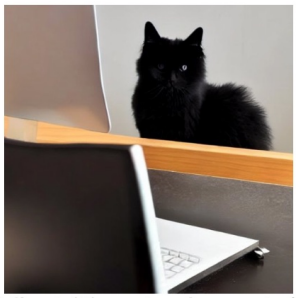
Examples

Q1. How well does the image match the description?

When you answer Q1, please judge only by the correspondence of content in the caption and the image. For example, we consider the caption and image below match exactly. Although the image has some irregular textures and shapes, all the important content, e.g., man, bike, and friend on the back, can be easily identified.

a man on his bike with his friend riding on the back

The image below shows all the objects described in the caption. However, the cat is not sitting on top of a keyboard. Please select Has a few minor discrepancies for an image with such a minor failure.



A black fluffy cat sitting on top of a computer keyboard.

Q2. Rate the quality of the image.

When you answer Q2, please rate if the image look like a real image. A high quality image is one that is hard to tell if it is real or AI-generated. Some AI-generated images may have artifacts such as irregular textures, shapes, or noise as in the examples below. Rate such images as lower quality.



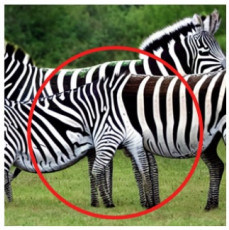


Figure 2. Instructions provided to the annotators.

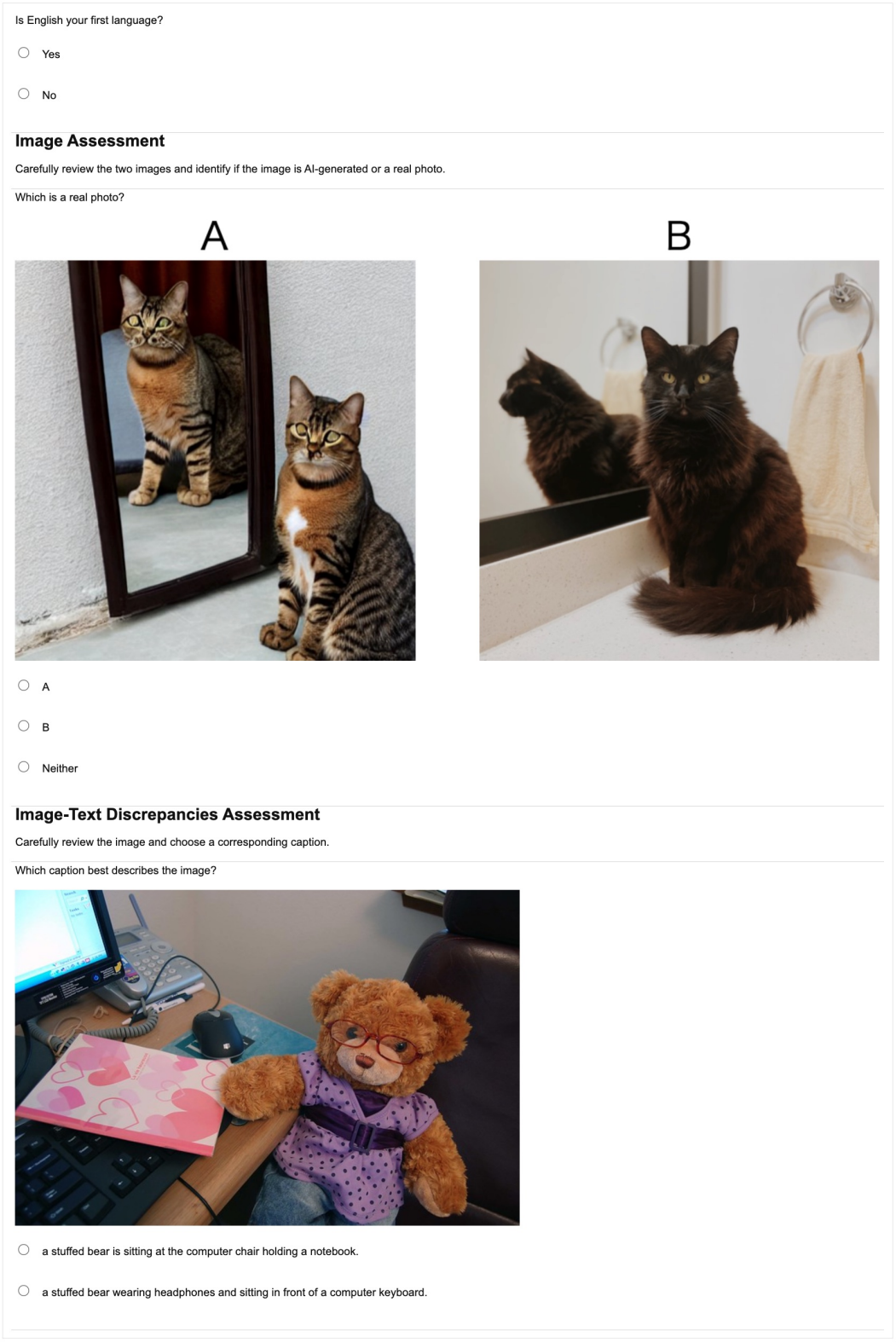


Figure 3. Screenshot of the pre-task qualification test for *skillfulness* qualification. Annotators who answer that their first language is English and give correct answers to the two quiz are allowed to work for our annotation task.

3. Detailed results of human and automatic evaluation

Figure 4 shows the human and automatic evaluation results on COCO. The result demonstrates that the automatic measures do not align with human evaluation. Table 3 shows the pairwise comparison results by a Tukey's HSD test. We also compute Hedge's g values which indicate the differences between two methods are over one standard deviation apart. We confirm that the ratings provided for each model show significant differences.

On DrawBench, Stable Diffusion and GLIDE obtain comparable ratings for fidelity as shown in Fig. 5. For alignment, annotators rate Stable Diffusion the best, while the other models do not show significant differences as in Tab. 4. On PartiPrompts, we observe similar trends, but the difference of fidelity ratings between Stable Diffusion and GLIDE is statistically significant as in Tab. 5.

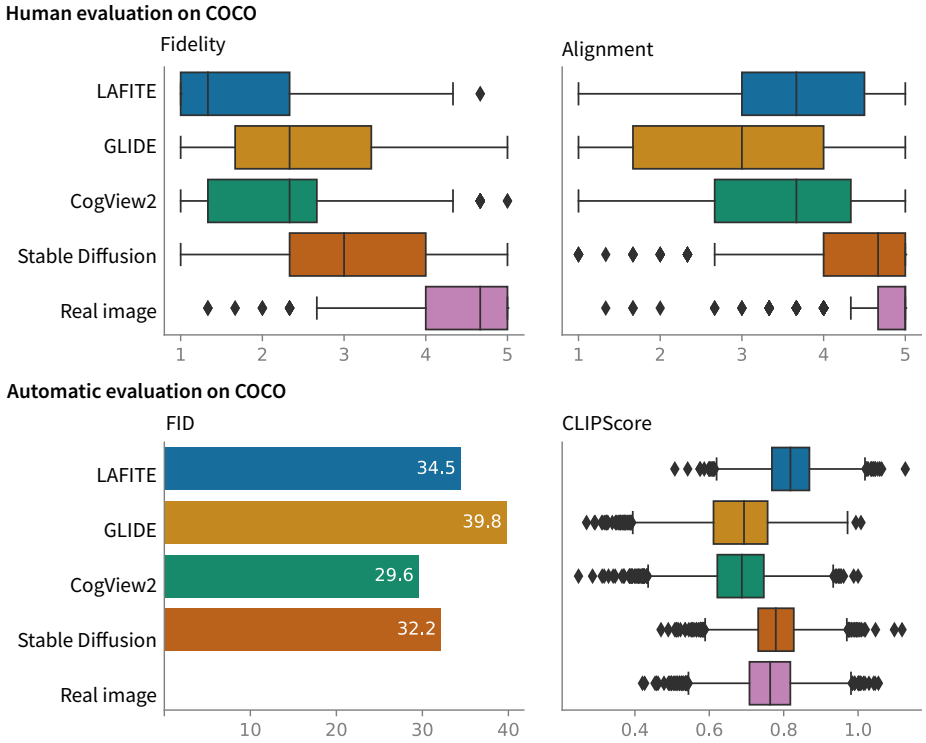


Figure 4. Evaluation results on COCO. (Top) Distributions of human ratings for fidelity and alignment. (Bottom) Automatic evaluation results. The bottom right plot shows the distribution of sample-level CLIPScore.

Table 3. Pairwise post-hoc test with Tukey's HSD test for ratings of Fidelity and Alignment on COCO. The numbers provided in the table are p-values, and the numbers in parentheses are effect size (Hedge's g).

	Real image	CogView2	GLIDE	LAFITE
Fidelity				
CogView2	0.0000 (-2.79)	—	—	—
GLIDE	0.0000 (-2.27)	0.0000 (0.39)	—	—
LAFITE	0.0000 (-3.69)	0.0000 (-0.48)	0.0000 (-0.89)	—
Stable Diffusion	0.0000 (-1.45)	0.0000 (0.84)	0.0000 (0.48)	0.0000 (1.31)
Alignment				
CogView2	0.0000 (-1.50)	—	—	—
GLIDE	0.0000 (-1.85)	0.0000 (-0.49)	—	—
LAFITE	0.0000 (-1.45)	0.0002 (0.18)	0.0000 (0.68)	—
Stable Diffusion	0.0000 (-0.67)	0.0000 (0.85)	0.0000 (1.28)	0.0000 (0.72)

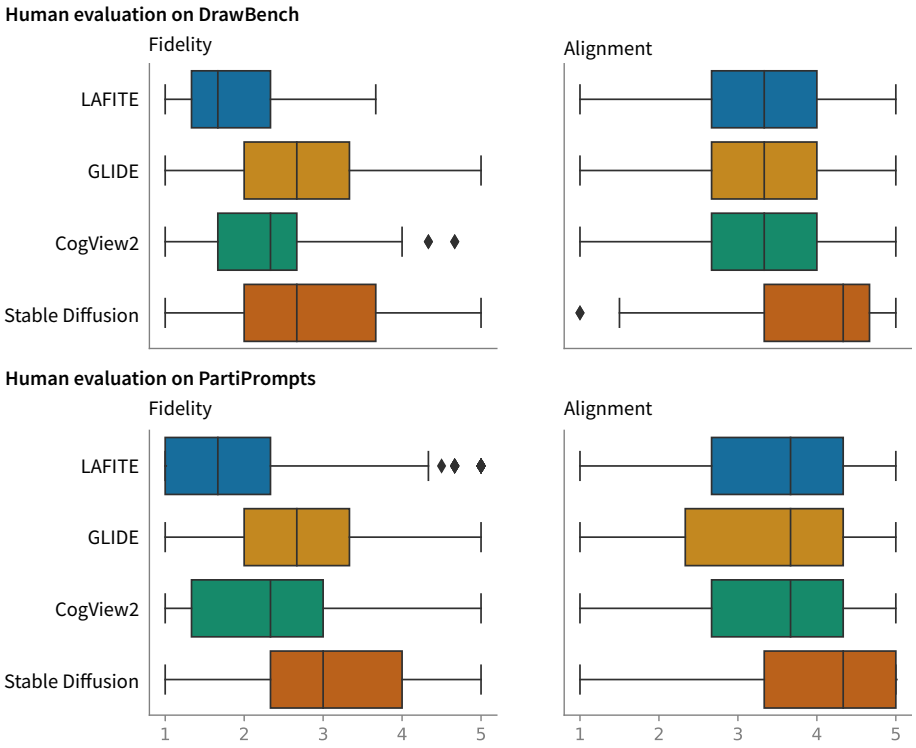


Figure 5. Distributions of human ratings for fidelity and alignment on DrawBench (top) and PartiPrompts (bottom).

Table 4. Pairwise post-hoc test with a Tukey's HSD test for ratings of Fidelity and Alignment on DrawBench. The numbers provided in the table are p-values, and the numbers in parentheses are effect size (Hedge's g).

	CogView2	GLIDE	LAFITE
Fidelity			
GLIDE	0.0000 (0.48)	—	—
LAFITE	0.0000 (-0.58)	0.0000 (-1.06)	—
Stable Diffusion	0.0000 (0.65)	0.1034 (0.21)	0.0000 (1.19)
Alignment			
GLIDE	0.5558 (0.13)	—	—
LAFITE	0.9352 (0.06)	0.8884 (-0.07)	—
Stable Diffusion	0.0000 (0.71)	0.0000 (0.64)	0.0000 (0.70)

Table 5. Pairwise post-hoc test with Tukey's HSD test for ratings of Fidelity and Alignment on PartiPrompts. The numbers provided in the table are p-values, and the numbers in parentheses are effect size (Hedge's g).

	CogView2	GLIDE	LAFITE
Fidelity			
GLIDE	0.0000 (0.34)	—	—
LAFITE	0.0000 (-0.37)	0.0000 (-0.73)	—
Stable Diffusion	0.0000 (0.67)	0.0000 (0.33)	0.0000 (1.07)
Alignment			
GLIDE	0.1371 (-0.07)	—	—
LAFITE	0.9504 (0.02)	0.0363 (0.10)	—
Stable Diffusion	0.0000 (0.57)	0.0000 (0.62)	0.0000 (0.58)

4. Captions and images used for sample size analysis

The caption and image pairs used for annotation size analysis are shown in Fig. 8. We selected samples where three annotators gave diverse labels in a pilot data collection. The histogram plots show distributions of human ratings by 60 annotators. We observe variations in the ratings, and some samples show several peaks. This observation indicates that human evaluation measures other than averaging ratings may be needed.

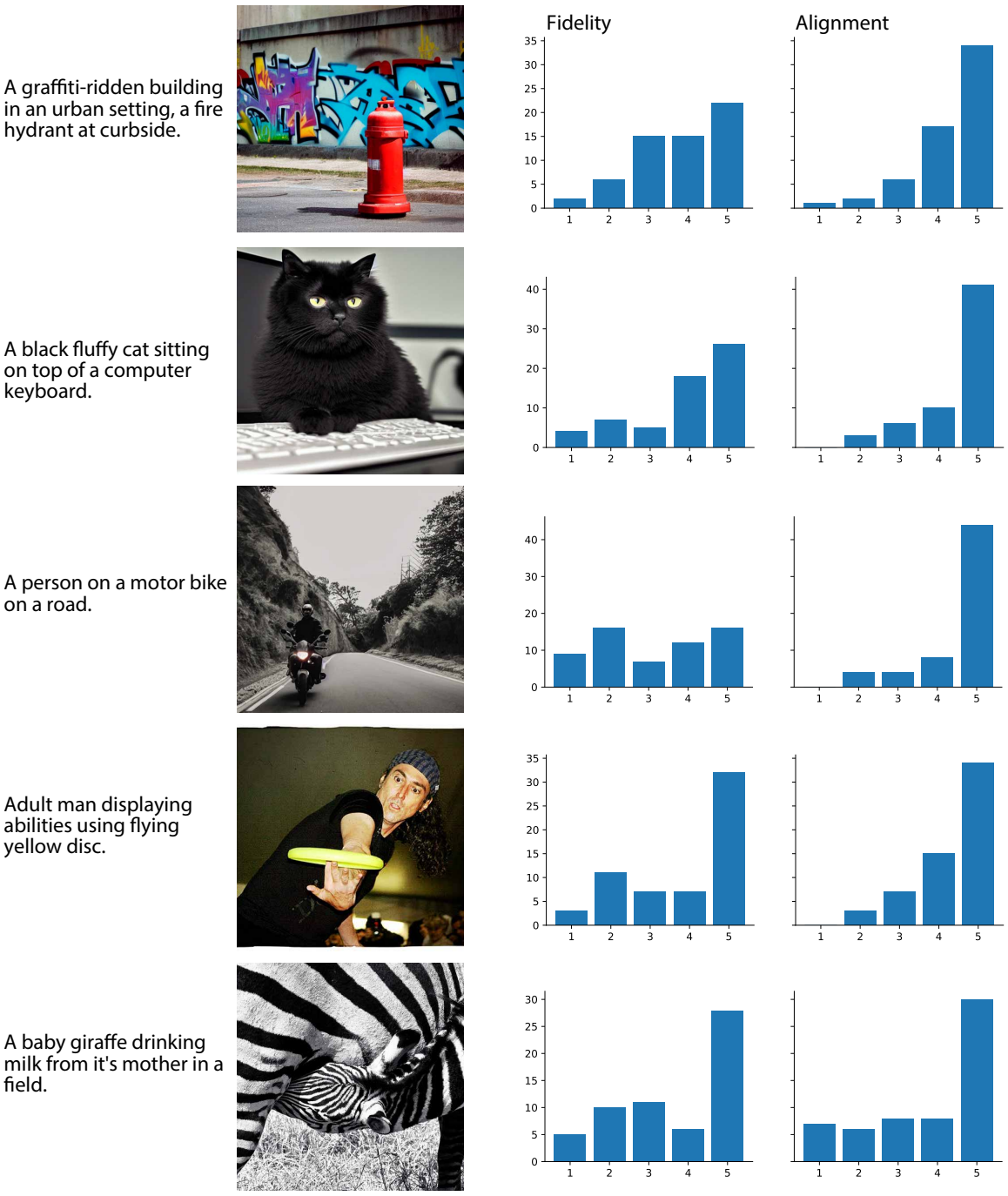


Figure 6. Image and caption pairs used for annotator size analysis. Histograms represent distributions of human ratings.

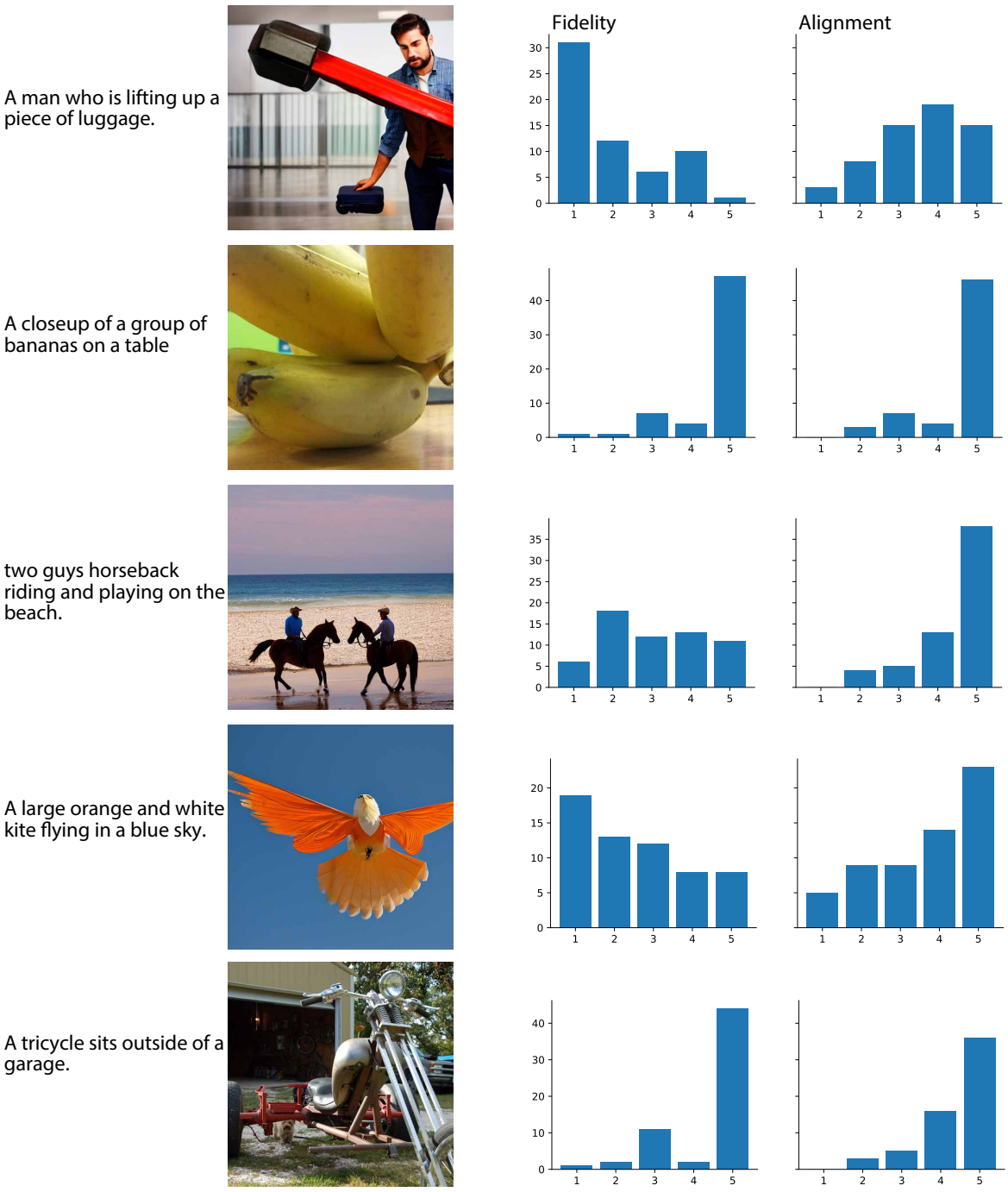


Figure 7. Image and caption pairs used for annotator size analysis. Histograms represent distributions of human ratings.

Figure 8. Image and caption pairs used for annotator size analysis. Histograms represent distributions of human ratings.

5. Reporting experimental details for transparency

Our literature review revealed that many papers omit details of experimental configurations of human evaluation. To the alleviate transparency issue, we offer templates for reporting human evaluation settings. Table 6 summarizes recommended details to report. For customizable sample text, see Figure 9.

Table 6. Example report of a human evaluation setting.

Dataset details	
#captions	1000
#ratings / item	3
#unique annotators	148
Tested models	LAFITE, GLIDE, CogView2, Stable Diffusion, Real image
Types of rating	5-point Likert scale
Evaluation criteria	Fidelity, Alignment to caption
Annotation details	
Platform	AMT
Annotator qualification	i) Over 18 years old and agreed to work with potentially offensive content. ii) AMT Masters
Compensation	$0.05 / task
Interface	Figure 1
Instructions	Figure 2
IAA	Fidelity: 0.41, Alignment: 0.48 (Krippendorff's α)

We collected annotations for images generated by [tested models] for [#captions] captions, resulting in [#annotations in total] annotations. Annotators are invited on [crowdsourcing platform]. Annotators who [annotator qualification] are allowed to work on our task. Krippendorff's α is [α value]. [#unique annotators] annotators participated in total, and the average number of tasks per annotator was [average #tasks per annotator]. Annotators get [compensation per task] for each instance of the task The median time spent on one task is [time spent per task]; that is, the expected hourly wage is [expected hourly wage].

Figure 9. Sample template for reporting human evaluation settings.

6. Potential Bias in Annotator Ratings

Even with carefully designed instructions for annotators, they may annotate a text-image pair differently. The causes of the disagreement between ratings of multiple annotators are, for example, the subjectivity of annotation tasks and the different stringency of annotators. We therefore describe a potential concern of annotator biases in datasets collected through crowdsourcing.

Each plot in Figure 10 shows the distribution of the mean ratings of the annotators in our collected dataset. The x- and y-axes indicate the mean rating score for each annotator and the density of annotators, respectively. The mean alignment/fidelity scores are computed by taking the average of one annotator's ratings for different samples. We observe a non-negligible number of annotators who provide highly biased ratings (the right and left tails of distributions). This result suggests that there may exist annotator-dependent rating biases. However, this can also happen when the tasks assigned to each annotator have imbalanced true ratings; in this case, annotators with biased mean ratings may correctly judge their tasks. We consider a rating correction strategy for task-dependent rating biases to remove the effect of unbalanced task assignments. We first compute the mean ratings for each task (*i.e.*, text-image pair) and then normalize each rating using the mean. Figure 11 demonstrates the distributions of mean corrected ratings for fidelity and alignment. It can be observed that the distributions are more "centered", and the density of annotators with extreme mean scores is reduced compared to that without the rating correction. On the other hand, there are a few annotators with biased (corrected) ratings, particularly in mean alignment scores.

In summary, the annotator-dependent bias in quality ratings is not severe in our collected dataset; this is also confirmed by the high Krippendorff's α values. We may further enhance the data quality by considering the annotator-dependent bias in the aggregation of multiple ratings to generate reliable ground truth labels.

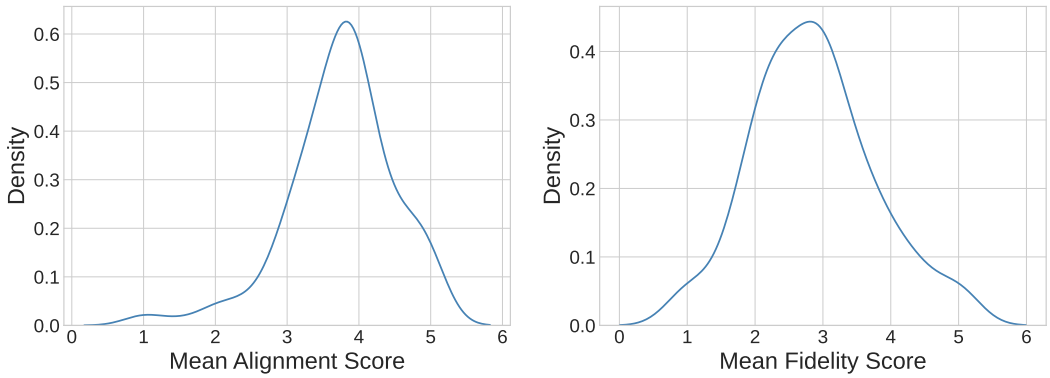


Figure 10. Distributions of mean ratings of each annotator for fidelity and alignment.

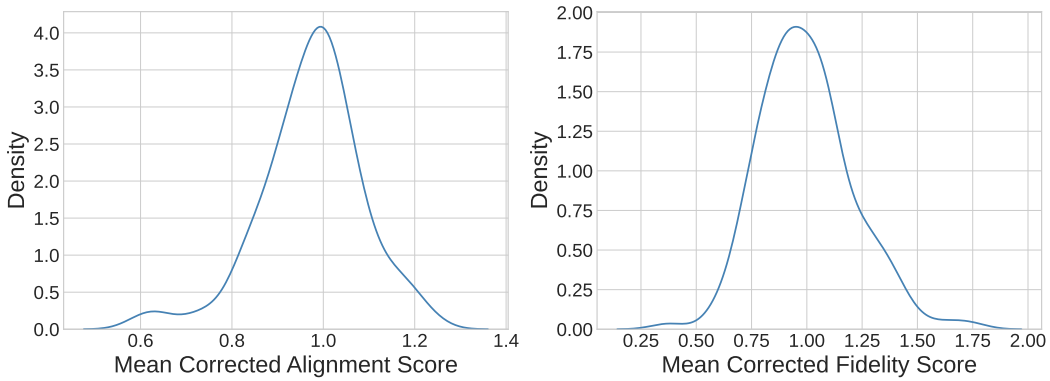


Figure 11. Distributions of mean corrected ratings of each annotator for fidelity and alignment.

References

[1] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. RiFeGAN: Rich Feature Generation for Text-to-Image Synthesis From Prior Knowledge. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10911–10920, 2020. 2

[2] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. In *Adv. Neural Inform. Process. Syst.*, volume 34, pages 19822–19835, 2021. 2

[3] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers, 2022. arXiv:2204.14217 [cs]. 2

[4] Tan M. Dinh, Rang Nguyen, and Binh-Son Hua. TISE: Bag of Metrics for Text-to-Image Synthesis Evaluation. In *Eur. Conf. Comput. Vis.*, volume 13696, pages 594–609, 2022. 2

[5] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Eur. Conf. Comput. Vis.*, page 19, 2022. 2

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, 2022. arXiv:2208.01618 [cs]. 2

[7] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10696–10706, 2022. 2

[8] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. In *CVPR*, pages 7986–7994, 2018. 2

[9] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-Based Real Image Editing with Diffusion Models, 2022. arXiv:2210.09276 [cs]. 2

[10] Qicheng Lao, Mohammad Havaei, Ahmad Pesaranghader, Francis Dutil, Lisa Di Jorio, and Thomas Fevens. Dual Adversarial Inference for Text-to-Image Synthesis. In *Int. Conf. Comput. Vis.*, pages 7567–7576, 2019. 2

[11] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable Text-to-Image Generation. In *Adv. Neural Inform. Process. Syst.*, volume 32, 2019. 2

[12] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-Driven Text-To-Image Synthesis via Adversarial Training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12174–12182, 2019. 2

[13] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. StyleT2I: Toward Compositional and High-Fidelity Text-to-Image Synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18197–18207, 2022. 2

[14] Jiadong Liang, Wenjie Pei, and Feng Lu. CPGAN: Content-Parsing Generative Adversarial Networks for Text-to-Image Synthesis. In *Eur. Conf. Comput. Vis.*, page 18, 2020. 2

[15] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to Image Generation With Semantic-Spatial Aware GAN. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18187–18196, 2022. 2

[16] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. In *Eur. Conf. Comput. Vis.*, page 18, 2022. 2

[17] 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 *International Conference on Machine Learning*, volume 162, pages 16784–16804, 2022. 2

[18] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. MirrorGAN: Learning Text-To-Image Generation by Redescription. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1505–1514, 2019. 2

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation, 2021. arXiv:2102.12092 [cs]. 2

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 2

[21] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. DAE-GAN: Dynamic Aspect-Aware GAN for Text-to-Image Synthesis. In *Int. Conf. Comput. Vis.*, pages 13960–13969, 2021. 2

[22] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, 2022. arXiv:2208.12242 [cs]. 2

[23] 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. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, 2022. 2

[24] Hongchen Tan, Xiuping Liu, Xin Li, Yi Zhang, and Baocai Yin. Semantics-Enhanced Adversarial Nets for Text-to-Image Synthesis. In *Int. Conf. Comput. Vis.*, pages 10501–10510, 2019. 2

[25] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16515–16525, 2022. 2

[26] Fuxiang Wu, Liu Liu, Fusheng Hao, Fengxiang He, and Jun Cheng. Text-to-Image Synthesis Based on Object-Guided Joint-Decoding Transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18113–18122, 2022. 2

[27] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1316–1324, 2018. 2

[28] Kun Yan, Lei Ji, Chenfei Wu, Ming Zhou, Nan Duan, and Shuai Ma. Trace Controlled Text to Image Generation. In *Eur. Conf. Comput. Vis.*, page 17, 2022. 2

[29] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics Disentangling for Text-To-Image Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2327–2336, 2019. 2

[30] 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. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, 2022. arXiv:2206.10789 [cs]. 2

[31] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 833–842, 2021. 2

[32] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5908–5916, Venice, 2017. 2

[33] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks, 2018. arXiv:1710.10916 [cs, stat]. 2

[34] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic Text-to-Image Synthesis With a Hierarchically-Nested Adversarial Network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6199–6208, 2018. 2

[35] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards Language-Free Training for Text-to-Image Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17907–17917, 2022. 2

[36] Bin Zhu and Chong-Wah Ngo. CookGAN: Causality Based Text-to-Image Synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5519–5527, 2020. 2

[37] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5802–5810, 2019. 2