2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01372

Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation

Mayu Otani[1] Riku Togashi[1] Yu Sawai[1] Ryosuke Ishigami[1] Yuta Nakashima[2]
Esa Rahtu[3] Janne Heikkilä[4] Shin'ichi Satoh[1]

[1]CyberAgent, Inc. [2]Osaka University [3]Tampere University [4]University of Oulu

Abstract

*Human evaluation is critical for validating the performance of text-to-image generative models, as this highly cognitive process requires deep comprehension of text and images. However, our survey of 37 recent papers reveals that many works rely solely on automatic measures (*e.g.*, FID) or perform poorly described human evaluations that are not reliable or repeatable. This paper proposes a standardized and well-defined human evaluation protocol to facilitate verifiable and reproducible human evaluation in future works. In our pilot data collection, we experimentally show that the current automatic measures are incompatible with human perception in evaluating the performance of the text-to-image generation results. Furthermore, we provide insights for designing human evaluation experiments reliably and conclusively. Finally, we make several resources publicly available to the community to facilitate easy and fast implementations.*

1. Introduction

Text-to-image synthesis has seen substantial development in recent years. Several new models have been introduced with remarkable results. The majority of the works validate their models using automatic measures, such as FID [13] and recently proposed CLIPScore [12], even though many papers point out problems with these measures. The most popular measure, FID, is criticized for misalignment with human perception [9]. For example, image resizing and compression hardly degrade the perceptual quality but induce high variations in the FID score [28], while CLIPScore can inflate for a model trained to optimize text-to-image alignment in the CLIP space [27].

This empirical evidence of the misalignment of the automatic measures motivates human evaluation of perceived quality. However, according to our study of 37 recent papers, the current practices in human evaluation face significant challenges in reliability and reproducibility. We mainly identified the following two problematic practices. Firstly, evaluation protocols vary significantly from one paper to another. For example, some employ relative evaluation by simultaneously showing annotators two or more samples for comparison, and others collect scores of individual samples based on certain criteria. Secondly, important details of experimental configurations and collected data are often omitted. For example, the number of annotators who rate each sample is mostly missing, and sometimes even the number of assessed samples is not reported. These inconsistencies in the evaluation protocol make future analysis almost impossible. We also find that recent papers do not analyze the quality of the collected human annotations. Therefore, we cannot assess how reliable the evaluation result is. It is also difficult to know which is a good way to evaluate text-to-image synthesis among various evaluation protocols.

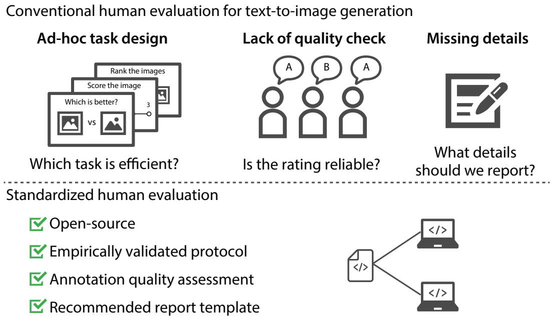

Conventional human evaluation for text-to-image generation
Ad-hoc task design
Rank the images
Score the image
Which is better?
vs
3
Which task is efficient?
Lack of quality check
A
B
A
Is the rating reliable?
Missing details
What details should we report?
Standardized human evaluation
Open-source
Empirically validated protocol
Annotation quality assessment
Recommended report template

Figure 1. Conventionally, researchers have used different protocols for human evaluation, and setup details are often unclear. We aim to build a standardized human evaluation.

The natural language generation (NLG) community has extensively explored human evaluation. Human evaluation is typically done in a crowdsourcing platform, and there are many well-known practices. Yet, quality control is an open challenge [16]. Recently, a platform for benchmarking multiple NLG tasks was launched [18]. The platform offers a

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01372

web form where researchers can submit their model predictions. The platform automatically enqueues a human evaluation task on AMT, allowing a fair comparison for its users.

We address the lack of a standardized evaluation protocol in text-to-image generation. To this end, we carefully design an evaluation protocol using crowdsourcing and empirically validate the protocol. We also provide recommendations for reporting a configuration and evaluation result.

We evaluate state-of-the-art generative models with our protocol and provide an in-depth analysis of collected human ratings. We also investigate automatic measures, *i.e.*, FID and CLIPScore, by checking the agreement between the measures and human evaluation. The results reveal the critical limitations of these two automatic measures.

Findings and resources Our main findings can be summarized as follows:

- **Reliability of prior human evaluation is questionable.** We provide insights about the required numbers of prompts and human ratings to be conclusive.
- **FID is inconsistent with human perception.** This is already known at least empirically, and our experiment supports it.
- **CLIPScore is already saturated.** State-of-the-art generative models are already on par with authentic images in terms of CLIPScores.

These findings motivate us to build a standardized protocol for human evaluation for better verifiability and reproducibility, facilitating to draw reliable conclusions. For continuous development, we open-source the following resources for the community.

- Implementation of human evaluation on a crowdsourcing platform, *i.e.*, Amazon Mechanical Turk (AMT)[1], which allows researchers to evaluate their generative models with a standardized protocol.
- Template for reporting human evaluation results. This template will enhance their transparency.
- Human ratings by our protocol on multiple datasets: MS-COCO [23], DrawBench [30], and PartiPrompts [37]. These per-image ratings, and not only their statistics, will facilitate designing automatic measures.

2. Related work

Human evaluation Evaluation of generative models, *e.g.*, for perceptual and linguistic data, inherently involves human perception and understanding, so human evaluation is inevitable. Nevertheless, there are no established evaluation practices, and researchers have been using different protocols [9, 26, 36–38]. To address challenges in human evaluation due to ad-hoc practices, prior studies offer shared human evaluation protocols, such as the one for unconditional image synthesis [39]. NLG community provides in-depth analysis of challenges in human evaluation on story generation [16]. To facilitate reliable model comparison based on human evaluation, a platform hosting human evaluation of multiple language generation tasks is proposed [18]. Inspired by these works, we aim to develop a shared evaluation protocol for text-to-image generation.

Some literature reviews have summarized challenges in crowdsourcing [2, 17]. They conclude that guaranteeing annotators' reliability is the main challenge. Annotators have a strong incentive to maximize their own monetary returns in the shortest possible period. As a result, they do not often pay enough attention to the tasks. The reviews offer actionable techniques to alleviate such problems.

Automatic evaluation The community has granted automatic measures, such as Inception Score (IS) [31], Frećhet Inception Distance (FID) [13], and Precision-Recall [20], as additional options for evaluation. IS evaluates whether generated images have identifiable objects and diversity using the output of the Inception model [34]. FID and Precision-Recall evaluate the discrepancy between distributions of real and generated images.

Evaluation of text-to-image generative models often leverages R-precision [35] and CLIPScore [12]. CLIPScore has recently been proposed to evaluate image-text alignment in image captioning, which is diverted to text-to-image generation. Concept detectors can also be used to assess if a predefined set of concepts in the prompt are detected in the generated image [24]. Meanwhile, many prior works pointed out the limitations of these automatic measures [3, 5, 7, 9, 25, 28]. Image processing operations, such as resizing and compression, often exemplify the misalignment problem of FID with human perception [25,28], which is empirically confirmed using human ratings [9].

3. Review: Human evaluation in text-to-image

We surveyed 37 recent text-to-image generation papers[2], and reviewed how they use and report human evaluation. The details of the surveyed papers are in the supplementary material. The following summarizes our findings.

Human evaluation vs. automatic measures Only 20 out of 37 papers provide human evaluation, which means that 17 works rely solely on automatic measures. Such measures are found to be inconsistent with human perception, therefore, should not be used as the only measure.

[1] https://www.mturk.com/

[2] We collected papers by querying "text to image" in CVF conferences, including CVPR, ICCV, and ECCV from 2017 to 2022. We also add recent successors. The full list of papers is in the supplementary material.

Number of samples and ratings Among 20 papers, 18 papers report the number of samples. However, only four disclose the number of ratings per sample even though the evaluation of generated images is highly subjective, and large discrepancies are expected in ratings. Moreover, crowdsourcing often suffers from noise and requires multiple annotations for each sample to be conclusive.

Evaluation criteria The overall quality of generated images and relevance to text prompts are major concerns in human evaluation; 18 papers assess overall quality and 14 papers assess relevance to text. Others include correctness of object locations [36] and consistency of multiple image generation [26]. This implies that some papers only evaluate a single aspect of generated images.

Rating methods We identify three different methods to collect ratings. Ten papers adopt a comparative approach by choosing the best among two or more samples, while nine works use comparative judgment but require ranking multiple samples. Three papers use a 5- or 3-point Likert scale for rating. There are wide discrepancies in the way of conducting human evaluation in different papers. However, their validity is rarely discussed.

Annotation quality assessment We find a problematic practice of not reporting the quality of annotations. A typical metric is an inter-annotator agreement (IAA), such as Cohen's κ and Krippendorff's α [19]. No paper reports IAA, which raises a concern about the reliability of results.

Sample size Many papers use less than 100 samples for each model for human evaluation [8–10, 21, 22]. Such a small sample size occasionally leads to different conclusions. Our experiment on COCO in Section 5.3 demonstrates that more than 500 samples are necessary for a stable ranking of competing models; otherwise, the ranking changes easily by chance with different samples.

Compensation and qualification Most papers do not reveal compensation for annotators and qualification filters, despite the fact that the current ethical standard recommends reporting basic statistics on time commitment and compensation [33].

Interface design The user interface design for annotation offers a high degree of freedom, and various choices can impact ratings. For instance, a constant image resolution of 256x256 was used in [9], while a mixture of resolutions for different models were applied in [14, 38]. Moreover, a real image may be displayed side-by-side for reference as in [9, 37]. Such details of the user interface are often undisclosed. While researchers often share their model code, it is less common for human evaluation interfaces. Designing instructions, tasks, and rating options is critical and requires substantial consideration. The lack of reusable resources hinders the continuous improvement of human evaluation protocols and practices.

4. Our design for text-to-image evaluation

We decided to use a crowdsourcing platform for human evaluation. The design of our evaluation task follows two principles. Firstly, *the task should be simple*. We make it so simple that even inexperienced annotators can finish an instance of the task without extra effort to be familiar with it. Secondly, *evaluation results should be interpretable*. Human evaluation is for helping researchers understand the models output; thus resulting annotation data should be useful for follow-up analysis. For example, the evaluation that produces relative rankings is difficult to interpret as each sample's rating depends on other samples. For better interpretability, direct scoring is preferred in recent human evaluations for natural language generation [6, 16, 18].

4.1. Rating format

There are two major options in rating methods: comparative and absolute. Comparative evaluation, such as ranking generated images, is usually easier for annotators, and their ratings tend to be consistent. However, comparative evaluation needs baseline models shared among all evaluation attempts. At least currently, generative models enjoy rapid advancement, which can make baselines outdated in a short period. Another problem is its limited interpretability. Comparative evaluation only tells relative goodness among a given set of images but does not care about the goodness among all. One image can win because either it is of high quality or the baseline is weak. Lastly, comparative evaluation results in a relative ranking of models, and thus diachronic comparison (or comparison over time) among different evaluation results is almost infeasible.

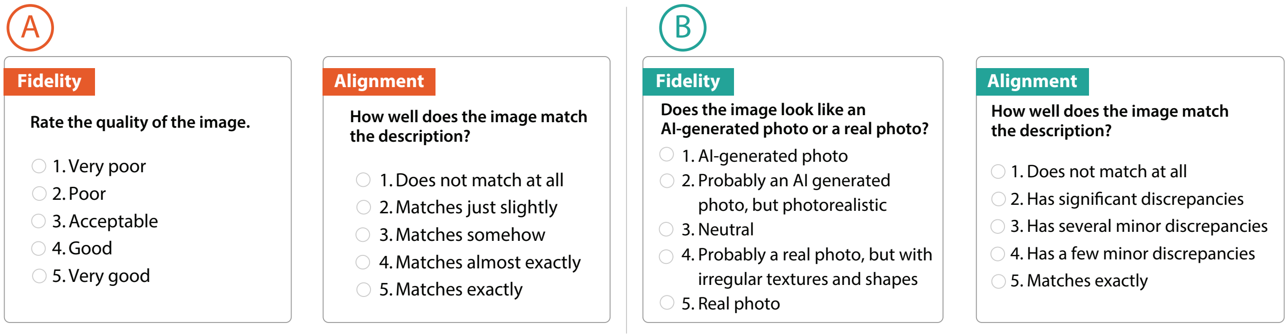
Considering these limitations, we decided to adopt absolute evaluation as our basic rating method. Yet absolute evaluation has some challenges. Firstly, it is harder than comparative evaluation and tends to result in more noisy annotations. Instructions, questions, and options (labels) must be carefully designed for quality control.

4.2. Evaluation criteria and wording

Our survey shows that many prior works employ *fidelity* and *alignment* as evaluation criteria. Fidelity means how well a generated image looks like a real photo, whereas alignment means how well a generated image matches the text. We consider that these two criteria represent sufficiently the essential aspects of the quality of text-to-image generation and decide to follow the convention.

The wording of questions and option labels can largely affect annotators' labeling behavior. A common indiscretion is to provide only endpoint labels, *e.g.*, Likert scales 1: *worst*, 5: *best*, and the other options are unlabeled.

In our pilot data collection experiment, we tried two design candidates (Fig. 2) to investigate the impact of con-



A

Fidelity

Rate the quality of the image.

○ 1. Very poor
○ 2. Poor
○ 3. Acceptable
○ 4. Good
○ 5. Very good

Alignment

How well does the image match the description?

○ 1. Does not match at all
○ 2. Matches just slightly
○ 3. Matches somehow
○ 4. Matches almost exactly
○ 5. Matches exactly

B

Fidelity

Does the image look like an AI-generated photo or a real photo?

○ 1. AI-generated photo
○ 2. Probably an AI generated photo, but photorealistic
○ 3. Neutral
○ 4. Probably a real photo, but with irregular textures and shapes
○ 5. Real photo

Alignment

How well does the image match the description?

○ 1. Does not match at all
○ 2. Has significant discrepancies
○ 3. Has several minor discrepancies
○ 4. Has a few minor discrepancies
○ 5. Matches exactly

Figure 2. Question and labels of two candidate task designs. A uses typical labels for a 5-point Likert scale. B's labels are more precise.

creteness in the questions and option labels. (A) is a baseline task whose wording follows typical Likert scale labels. (B) describes questions and options more specifically. For alignment, (B) uses more detailed quantifiers than (A) to reduce subjectivity. For fidelity, (A) asks general quality and uses general labels, whereas (B) asks to judge if the image looks AI-generated or real. A similar question was employed in [39] that asks binary judgment to distinguish real images from generated ones.

For the pilot data collection, we sampled human annotations for 200 text-image pairs of COCO Captions [23]. Among these 200 samples, 100 images are generated by Stable Diffusion [29] conditioned by the captions, while the other 100 images are real ones. The annotators get paid \$0.04 for each evaluation task. We restrict the number of annotations per annotator to at most 40 to avoid a small set of annotators dominating all evaluation task instances. We screen annotators with maturity, experience, and location qualification filters explained in Section 4.3. As a result, Krippendorff's α for the alignment and fidelity questions in the case of (A) are 0.07 and 0.18, respectively, indicating high variations among annotators. On the other hand, (B) achieved 0.39 for fidelity and 0.26 for alignment. Although not surprising, these results successfully confirm that more specific questions and labels lead to significantly better IAA. The following experiments use the design (B).

4.3. Qualifications

Screening annotators is mandatory in AMT as the annotator pool is diverse and global. Some annotators do not have sufficient skills. For example, language fluency is critical for our task since annotators are required to align English text and image. AMT provides qualification filters that allow its users to employ annotators who meet task-specific needs. Meanwhile, it is not trivial to identify the sufficient and necessary set of qualification filters (too many filters may reduce the number of potential annotators). We thus experimentally explore qualification settings.

We tested the following qualification filters.

i) *Maturity*: Over 18 years old and agreed to work with potentially offensive content.
ii) *Experience*: Completed more than 5000 HITs with an approval rate larger than or equal to 99%.
iii) *Location*: Located in an English-speaking country[3].
iv) *Skillfulness*: Passed a pre-task qualification test with three simple questions to confirm basic skills to assess image quality and to align text and image.[4]
v) *Master*: Good-performing and granted AMT Masters.

As we are not fully aware of the content in generated images, we always use the maturity qualification. Although location qualification iii) is recommended in prior work in NLG [16, 18], VPN can easily cheat this qualification and is a common practice among annotators [17]. Also, it should be noted that a substantial group of residents in the US do not use English as their primary language.

To ablate the impact of qualification filters, we conduct pilot data collection whose configuration is the same as the one in Section 4. Three annotators gave ratings for each sample based on the questions and labels of (B) in Section 4.2. For ablation, we published the task on the same day but with different combinations of qualification filters.

Table 1 summarizes the results. We observed that the annotator group with maturity and experience qualifications spent the shortest time per instance. This may suggest inattention of the group, reflected in much lower IAA and much higher fidelity scores to generated images than the other groups. Adding location and skillfulness qualification filters shows positive impacts in terms of IAA, and annotators take more time to complete tasks. However, requiring more qualifications, especially skillfulness qualification, overly downsizes the annotator pool and prolongs the time to collect all annotations. AMT Masters group by itself achieved

[3] Following [16], we select annotators in US, Canada, UK, Australia, and New Zealand

[4] Examples of the qualification test are in the supplementary material.

[5] A positive IAA value indicates that the ratings are more consistent than random annotations. For example, the coherence rating of NLG in [16] achieves an IAA of 0.14.

Table 1. Comparison of four combinations of qualification filters. Scores for fidelity and alignment are computed by first taking the mean over all three human ratings for each sample and then taking the mean over all samples. We compute Krippendorff's α as an IAA measure[5]. Med. time is the median time between successive submissions as a proxy for time to complete a single instance of the task.

Qualification					Annotator performance			Stable Diffusion		Real image	
i	ii	iii	iv	v	Fidelity IAA	Alignment IAA	Med. Time	Fidelity	Alignment	Fidelity	Alignment
✓	✓				0.11	0.10	12.0	3.81	4.63	4.78	4.94
✓	✓	✓			0.39	0.26	16.0	2.83	4.18	4.43	4.76
✓	✓	✓	✓		0.37	0.40	20.0	2.71	4.23	4.32	4.67
✓				✓	0.53	0.44	25.0	2.65	4.18	4.58	4.81

a relatively high IAA and did not degrade the total annotation time. Therefore, we use the master qualification filter in the following experiments.

5. Human evaluation of existing models

We evaluate four text-to-image generative models in the zero-shot setting. We do not use prompt engineering or sampling for screening generated images.

Lafite [40] is a GAN model with CLIP text encoding. A model trained on Google CC3M [32] is used[6].

GLIDE [27] is a text-to-image diffusion model. We used the publicly-available lightweight variant[7]. We follow the authors' parameter settings.

CogView2 [9] is a hierarchical transformer-based model trained on a large-scale corpus of images with Chinese and English captions[8].

Stable Diffusion [29] is another diffusion model. We use the v1.4 model hosted by Hugging Face[9].

We use as sources of captions and images the datasets that are widely used in the literature; COCO Captions [23], DrawBench [30] and PartiPrompts [37]. COCO Captions provides images and five annotated captions for each image. We randomly pick one out of five as a prompt to generate an image. We discard invalid captions, such as "unable to see this image in this particular hit." DrawBench and PartiPrompts are prompt datasets tailored for benchmarking text-to-image generative models.

We follow the setting in Section 4, although annotators get $0.05 for each instance of the task and the limitation of annotations per annotator was increased to 250.

We summarize human ratings by first taking the mean over all human ratings for each sample and then averaging over all samples. Our fidelity evaluation is interpreted as a Mean Opinion Score (MOS) test [1] for visual quality, whereas alignment one is another test but similar to MOS.

5.1. Evaluation results

On COCO Captions, we collected annotations for images generated by the four models for 1,000 captions and real images of COCO Captions, resulting in 15,000 annotations. Krippendorff's α of the fidelity and alignment ratings are 0.41 and 0.48, respectively. 148 annotators participated in total, and the average number of tasks per annotator was 101.4. The median time spent on one task is 18 seconds; that is, the expected hourly wage is $10, which is compatible with an ethical recommendation [33]. Collecting 15,000 annotations took 30 hours. Examples of annotated caption-image pairs are shown in Fig. 3.

Human evaluation results on COCO Captions are summarized in Table 2. Real images in this dataset are preferred by human annotators in terms of both fidelity and alignment, and Stable Diffusion is the second best.

DrawBench provides 200 prompts. We generated images for all, resulting in 2,400 annotations. Krippendorff's α for fidelity and alignment are 0.13 and 0.19. The drop of IAA compared with COCO Captions can be due to the increase of difficulty in evaluation. DrawBench is a collection of challenging textual prompts including long text, rare words, misspellings *etc*. Models often fail for such prompts, and annotators experience difficulties in evaluating significant failures and complex text. 40 annotators participated in total, and the average number of instances per annotator is 60. The median time spent on a single instance is 14 seconds, so the expected hourly wage is $12.9. The overall time required to complete annotations was 1.7 hours.

On PartiPrompts, we collected annotations for images generated by the four models for 1,337 captions, resulting in 16,044 annotations. Krippendorff's α of the fidelity and alignment ratings are 0.21 and 0.40, respectively. 181 annotators participated in total, and the average number of tasks per annotator was 87.2. The median time spent on one task is 18 seconds. Collecting annotations took 48 hours.

The results on DrawBench and PartiPrompt are in Table 3. The model ranking in terms of fidelity is the same as COCO Captions. Stable Diffusion significantly outperform other models on both DrawBench and PartiPrompt in terms

[6] https://github.com/drboog/Lafite

[7] https://github.com/openai/glide-text2im

[8] https://github.com/THUDM/CogView2

[9] https://huggingface.co/CompVis/stable-diffusion-v1-4

Table 2. Human and automatic evaluation of generated and real images on MS-COCO. Rankings by human evaluation and automatic evaluation are misaligned.

model	Human		Automatic	
	Fidelity ↑	Alignment ↑	FID ↓	CLIPScore ↑
LAFITE [40]	1.77	3.73	34.46	0.82
GLIDE [27]	2.56	2.96	39.80	0.68
CogView2 [9]	2.19	3.55	29.57	0.68
Stable Diffusion [29]	3.09	4.35	32.19	0.78
Real Image	4.49	4.78	—	0.76

Table 3. Human evaluation results on DrawBench [30] and PartiPrompts [37].

	Fidelity	Alignment
DRAWBENCH		
LAFITE [40]	1.82	3.35
GLIDE [27]	2.67	3.41
CogView2 [9]	2.27	3.29
Stable Diffusion [29]	2.87	3.99
PARTIPROMPT		
LAFITE [40]	1.93	3.48
GLIDE [27]	2.71	3.38
CogView2 [9]	2.34	3.49
Stable Diffusion [29]	3.05	4.07

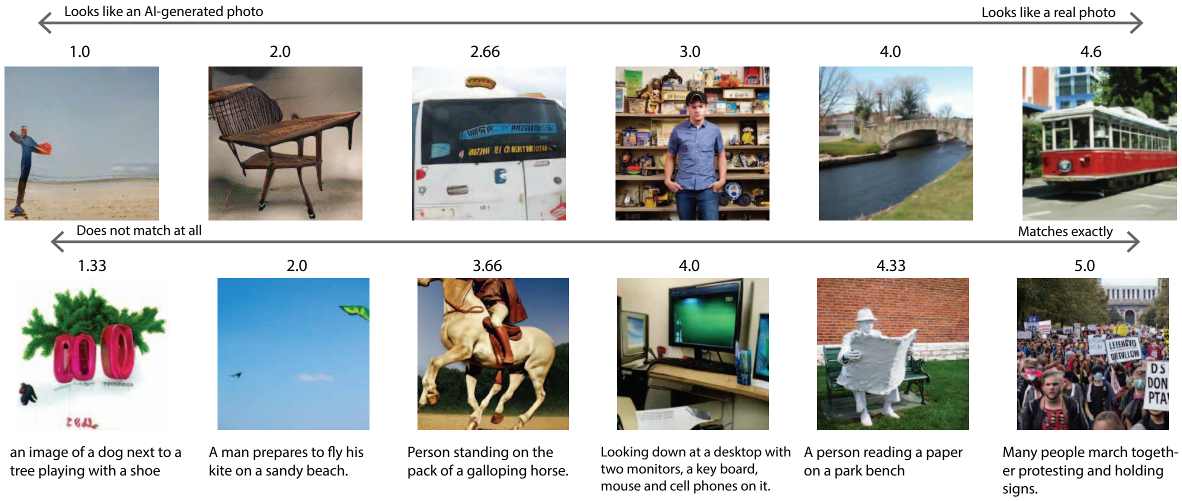

Looks like an AI-generated photo
Looks like a real photo
1.0
2.0
2.66
3.0
4.0
4.6
Does not match at all
Matches exactly
1.33
2.0
3.66
4.0
4.33
5.0
an image of a dog next to a tree playing with a shoe
A man prepares to fly his kite on a sandy beach.
Person standing on the pack of a galloping horse.
Looking down at a desktop with two monitors, a key board, mouse and cell phones on it.
A person reading a paper on a park bench
Many people march together protesting and holding signs.

Figure 3. Generated or real images and their human ratings of fidelity and alignment. The scores are the mean of three annotators' ratings.

of alignment. On the other hand, other models did not show statistically significant difference. More detailed results are in the supplementary material.

5.2. Agreement between automatic measures and human evaluation

Fréchet Inception Distance Table 2 shows FID values. Many prior works reported that FID does not align with perceived quality [9,25,28], and our human evaluation supports their claim. FID ranks CogView2 the best, while human annotators rated Stable Diffusion the best among the models in terms of fidelity and CogView2 in the third place. Figure 4 shows generated samples by the models. We observe that CogView2 produces more artifacts than Stable Diffusion (Fig. 4). FID fails to capture such irregular textures. The ranks of LAFITE and GLIDE are also inconsistent between FID and human evaluation. These results suggest that validation of fidelity solely relying on FID can lead to inconsistency with human perception.

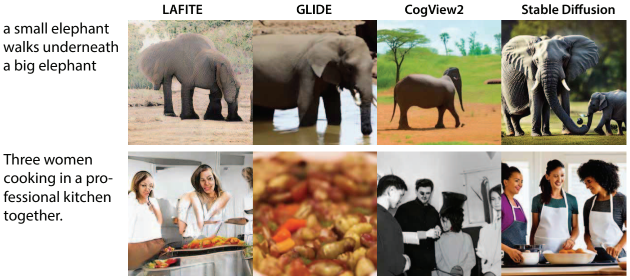

LAFITE
GLIDE
CogView2
Stable Diffusion
a small elephant walks underneath a big elephant
Three women cooking in a professional kitchen together.

Figure 4. Examples of input captions and generated images.

CLIPScore We investigate CLIPScore [12], a recently proposed automatic measure for text-to-image alignment.

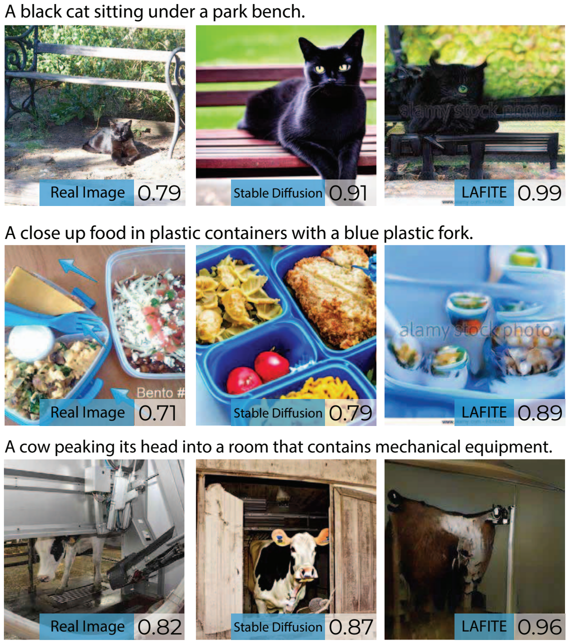

A black cat sitting under a park bench.
Real Image 0.79
Stable Diffusion 0.91
LAFITE 0.99
A close up food in plastic containers with a blue plastic fork.
Real Image 0.71
Stable Diffusion 0.79
LAFITE 0.89
A cow peaking its head into a room that contains mechanical equipment.
Real Image 0.82
Stable Diffusion 0.87
LAFITE 0.96

Figure 5. Below each caption, a real image and two generated images using the caption are displayed. Their sample-level CLIPScore is displayed in the bottom right of each. LAFITE achieves high scores, which are counter-intuitive.

Interestingly, LAFITE and Stable Diffusion achieve higher scores than image and caption pairs from COCO Captions, while human annotators rate the real images the best in terms of alignment as shown in Table 2. However, as observed in Figure 5, better CLIPScore does not necessarily mean better alignment, and the scale of the scores does not represent how well the image and text are aligned.

LAFITE's score is substantially higher than others. This may be because LAFITE involves optimization with respect to the CLIP-based GAN loss as discussed in [27]. On the other hand, Stable Diffusion gives higher scores without optimization in the CLIP space. However, CLIPScore does not discriminate the alignment performance between real images and Stable Diffusion generated images, while human annotators rate real images higher.

The comparison to human evaluation and qualitative result clarifies the limitation of CLIPScore. Evaluation measures are often the objectives in optimization [4]. Such optimization, however, suffers from adversarial effects. LAFITE, which optimizes CLIPScore, seems to exhibit this problem. Another problem is that CLIPScore cannot discriminate real images from images generated by Stable Diffusion, despite human annotators being able to distinguish them. This suggests that even with authentic caption-image pairs, there may be a gap in the CLIP space, which is not surprising, and images generated by Stable Diffusion are already within this margin. That is, CLIPScore is already saturated and may no longer be useful to evaluate state-of-the-art generative models.

5.3. Effect of sample size

The sample size is a major factor in experimental design. That is, the reliability of conclusions depends much on the number of human ratings, while it is directly reflected in monetary costs. The sample size here involves (1) the number of prompts and (2) the number of annotators who evaluate the same image.

Number of prompts We repeatedly computed the fidelity and alignment scores 500 times over n samples randomly drawn from 1000 COCO Captions. Figure 6 shows the mean over the 500 trials with the 5%-95% percentile interval. For both fidelity and alignment scores, there are large overlaps of the intervals when n is small. This suggests that conclusions drawn from small n may be unstable and easily flipped depending on the choice of prompts. With our choice of four models, we need more than 100 prompts to obtain consistent conclusions. Figure 6 gives a useful insight into the relationship between the difference in scores and the conclusive number of prompts.

Number of annotators for a single image We selected 13 captions from COCO Captions as prompts[10] and collected 60 human ratings for each of the corresponding 13 real images and images generated by Stable Diffusion. We randomly sampled m ratings out of 60 and computed the scores 500 times. This time, we computed the gain from the real images and Stable Diffusion images for each trial[11]. Figure 7 shows their mean and 5%-95% percentile interval for m. There are similar trends in both fidelity and alignment: Evaluation with fewer ratings often leads to instability. Particularly for alignment, we can occasionally draw a conclusion of the superiority of Stable Diffusion to real images. In the scenario of the lack of per-sample annotations, we may avoid unreliable conclusions by reporting (1) the statistical significance and (2) the effect size (*e.g.*, Hedge's g [11]). These reliability checks are recommended especially for low-budget experiments. Taking into account the deviation of the scores, ensuring a gain greater than 0.5 may be an easy way to conservative conclusions.

6. Discussions

6.1. Recommendations for crowdsourcing

The comparison between automatic measures and human evaluation reveals that current automatic measures are insufficient to represent human perception, and relying solely on them risks the reliability of conclusions. A careful discussion is recommended based on both automatic mea-

[10] The selection strategy is described in the supplementary material.

[11] The gains are mostly negative as real images' scores are mostly better.

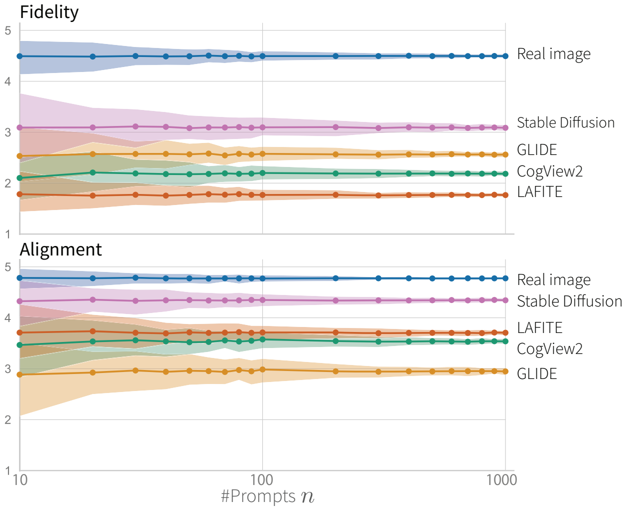


Figure 6. Effect of the number of prompts in the evaluation dataset. We compute the fidelity and alignment scores over sampled n prompts. Each data point represents the mean over the 500 trials, and the colored area represents the 5%-95% percentile interval. With a small number of test prompts, human evaluation can produce different rankings of the models by chance.

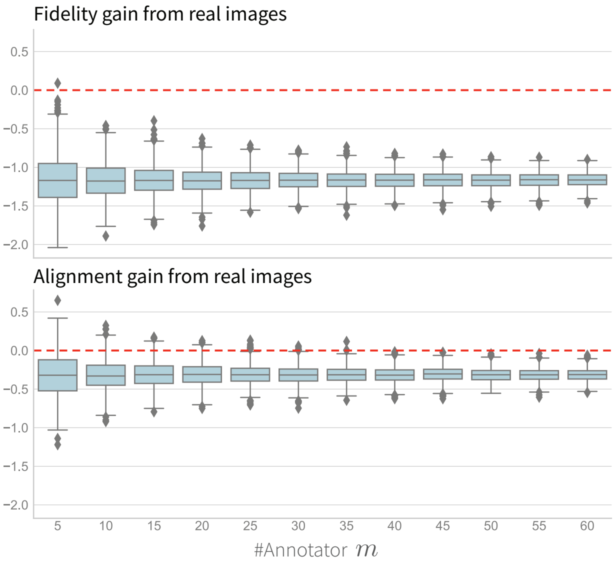


Figure 7. Effect of the number of raters per sample. The gain below the red dashed line indicates that real images outperform Stable Diffusion. The positive gains with few ratings demonstrate that few ratings per sample lead to instability.

sures and human evaluation, of which protocol is not yet matured. We thus propose a guideline for better evaluation.

Reporting experimental details for transparency Due to the difficulty of controlling annotation quality, results can vary over different runs of the same annotation process [16]. Literature should provide detailed description of the annotation for verifiability and reproducibility. Based on the recommendations for reporting experiments using crowdsourcing [2], we offer a sample template for reporting human evaluation settings in the supplementary material.

Understanding crowdworkers Monitoring annotators' behavior is an essential way to quality control. Using automation tools for efficiently completing many tasks is a common practice for crowdworkers [15], but these tools cause unnatural submission logs. For example, we observed many submissions that supposedly come from automating task approval. There are many other tools, *e.g.*, for annotation interface optimization, time management, and displaying requesters' reputations. Tasks that are incompatible with such tools limit annotator pools.

6.2. Limitations and future work

Our human evaluation protocol considers fidelity and alignment as two important criteria. However, there are other important criteria to consider, such as undesired bias in generated images, which is crucial in various applications. It is also important to evaluate this aspect.

We focus on natural images, but text-to-image generation can also create artwork. Different domains may require different evaluation criteria, *e.g.*, likability, and aesthetics.

AMT master qualification has limitations: i) it is limited to AMT, and ii) criteria of master qualification is unclear. Combining other qualifications substitutes for AMT master qualification to some extent. Post-hoc filtering of annotators based on IAA will also improve annotation quality.

A critical challenge in human evaluation is that collecting annotations large enough to get a reliable result is still expensive. One plausible remedy is to involve sampling techniques to effectively estimate models' performance with a smaller number of samples.

7. Conclusion

Our survey on human evaluations in recent text-to-image literature reveals reliability and transparency issues in human evaluation. We thus develop a human evaluation protocol for text-to-image generation. Our experiments demonstrated that automated measures do not align with human perception and are already getting outdated. The community needs to keep updating the automatic measures to catch up with the evolution of generative models. Yet, human evaluation itself is a challenging problem, and our design may not be optimal. We share our code with the community for continuous improvement in efficiency and coverage.

Acknowledgement This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, JSPS KAKENHI Grant-in-Aid for Scientific Research (A), the Academy of Finland projects 324346 and 353139.

References

[1] Recommendation ITU-T P.800, 1996. 5

[2] Herman Aguinis, Isabel Villamor, and Ravi S. Ramani. Mturk research: Review and recommendations. *Journal of Management*, 47(4):823–837, 2021. 2, 8

[3] Ahmed Alaa, Boris van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *ICML*, volume 162, pages 290–306, 2022. 2

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 7

[5] Ching-Yuan Bai, Hsuan-Tien Lin, Colin Raffel, and Wendy Chi-wen Kan. On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 2534–2542, 2021. 2

[6] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In *Conference on Machine Translation*, pages 1–55, 2020. 3

[7] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 2

[8] 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 *NeurIPS*, volume 34, pages 19822–19835, 2021. 3

[9] 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]. 1, 2, 3, 5, 6

[10] Tan M. Dinh, Rang Nguyen, and Binh-Son Hua. TISE: Bag of metrics for text-to-image synthesis evaluation. In *ECCV*, volume 13696, pages 594–609, 2022. 3

[11] Larry V Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981. 7

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528, 2021. 1, 2, 6

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, page 6629–6640, 2017. 1, 2

[14] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. In *CVPR*, pages 7986–7994, 2018. 3

[15] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey Bigham. Striving to earn more: A survey of work strategies and tool use among crowd workers. *AAAI Conference on Human Computation and Crowdsourcing*, 6(1):70–78, 2018. 8

[16] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using Mechanical Turk to evaluate open-ended text generation. In *EMNLP*, pages 1265–1285, 2021. 1, 2, 3, 4, 8

[17] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020. 2, 4

[18] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. GENIE: A leaderboard for human-in-the-loop evaluation of text generation. In *EMNLP*, 2022. 1, 2, 3, 4

[19] Krippendorff Klaus. Content analysis: An introduction to its methodology, 1980. 3

[20] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019. 2

[21] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. StyleT2I: Toward compositional and high-fidelity text-to-image synthesis. In *CVPR*, pages 18197–18207, 2022. 3

[22] Jiadong Liang, Wenjie Pei, and Feng Lu. CPGAN: Content-parsing generative adversarial networks for text-to-image synthesis. In *ECCV*, page 18, 2020. 3

[23] 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 *ECCV*, pages 740–755, 2014. 2, 4, 5

[24] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2

[25] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. An improved evaluation framework for generative adversarial networks. *arXiv preprint arXiv:1803.07474*, 2018. 2, 6

[26] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. StoryDALL-E: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*, page 18, 2022. 2, 3

[27] 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 *ICML*, volume 162, pages 16784–16804, 2022. 1, 5, 6, 7

[28] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *CVPR*, pages 11410–11420, 2022. 1, 2, 6

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 4, 5, 6

[30] 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. *arXiv preprint arXiv:2205.11487*, 2022. 2, 5, 6

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, page 2234–2242, 2016. 2

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 5

[33] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. Responsible research with crowds: Pay crowdworkers at least minimum wage. *Commun. ACM*, 61(3):39–41, 2018. 3, 5

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 2

[35] 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 *CVPR*, 2018. 2

[36] Kun Yan, Lei Ji, Chenfei Wu, Ming Zhou, Nan Duan, and Shuai Ma. Trace controlled text to image generation. In *ECCV*, page 17, 2022. 2, 3

[37] 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, 3, 5, 6

[38] 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 *ICCV*, pages 5908–5916, 2017. 2, 3

[39] Sharon Zhou, Mitchell L. Gordon, Ranjay Krishna, Austin Narcomey, Li Fei-Fei, and Michael S. Bernstein. HYPE: A benchmark for human eye perceptual evaluation of generative models. In *NeurIPS*, pages 3444–3456, 2019. 2, 4

[40] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. LAFITE: Towards language-free training for text-to-image generation. In *CVPR*, 2022. 5, 6