Structured Human Assessment of Text-to-Image Generative Models

Ciprian A. Corneanu
Amazon
300 Boren Ave N, Seattle, WA 98109
cicorn@amazon.com

Qianli Feng
Amazon
300 Boren Ave N, Seattle, WA 98109
fengq@amazon.com

Aleix M. Martinez
Amazon
300 Boren Ave N, Seattle, WA 98109
maleix@amazon.com

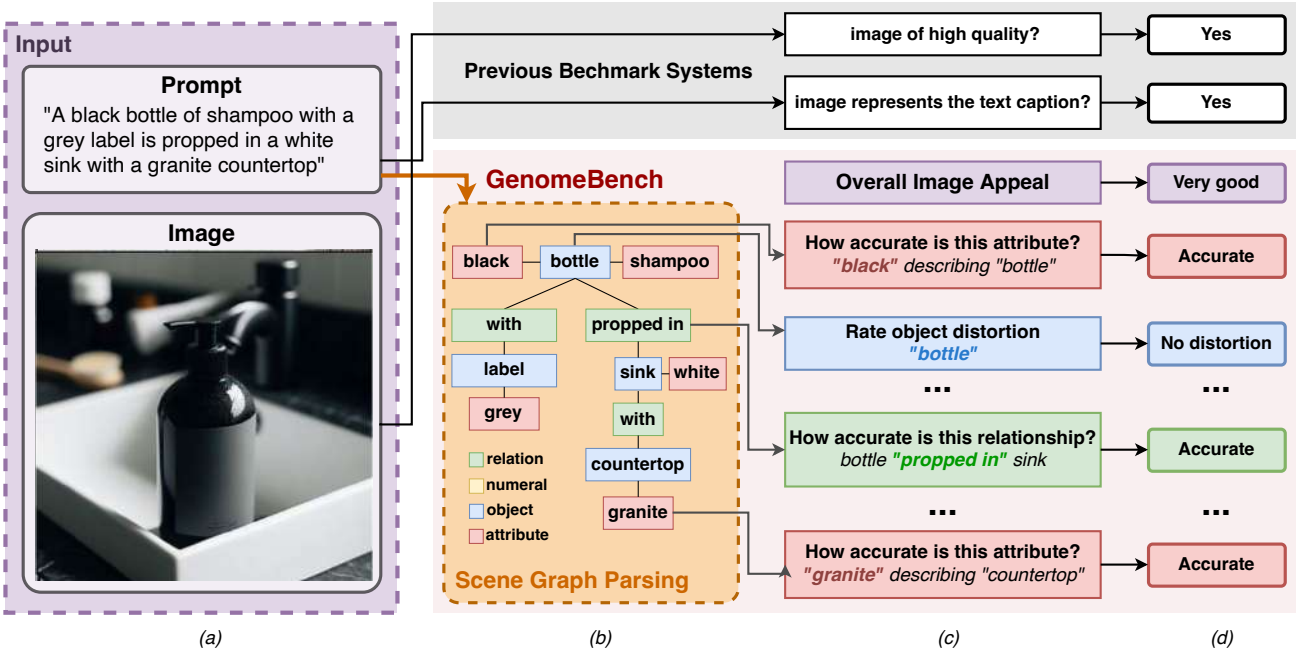

Figure 1. We propose GenomeBench, a novel framework to assess performance of text-to-image generative models. Given the prompt and the synthesized image (a), we parse the prompt into a scene graph (b) and ask humans carefully guided assessment questions (c). The answers are then summarized into a transparent score assessing text-to-image quality alignment (d). Comparing to simple, holistic questions asked in previous benchmarks, our method reduces annotation ambiguity and provides more insight into model performance.

Abstract

Following the great progress in text-conditioned image generation there is a dire need for establishing clear comparison benchmarks. Unfortunately, assessing performance of such models is highly subjective and notoriously difficult. Current automatic assessment of generated images quality and their alignment to text are approximate at best while human assessment is subjective, poorly calibrated and not very well defined. To address these concerns, we propose GenomeBench, a new framework for assessing quality of text-to-image generative models. It consists of a prompt dataset richly annotated with semantic components based on a formalized grounding of language and images. On top of it, we define a procedure to collect human assessment through a carefully guided question answering process. Finally, these assessments are summarized into a novel score built around quality and alignment to text. We show the

proposal achieves higher inter-annotator agreement with respect to the baseline human assessment and better correlation between quality and alignment compared to automatic assessment. Finally, we use this framework to dissect the performance of recent text-to-image models, providing insights on strength and weakness of each.

1. Introduction

Latest text-to-image generative models are very good at generating impressive images with remarkable diversity. Since their break-through, many new models have been proposed, each claiming some advantage over the state-of-the-art. Unfortunately, such comparisons are difficult to make because the relationship between text and images in current text-to-image generative models remains poorly defined.

To systematically perform such comparison, a variety of benchmarking systems have been developed [6, 13, 20, 25]. Typically, they consist of three major components: 1) a set of prompts for model input, 2) a set of questions probing annotators' perceptions, 3) a method for analyzing annotation results. Additionally, automatic metrics that strive to align with human perception have been proposed, eliminating the need of collecting costly human annotations at the cost of accuracy.

There are many desirable aspects of a generative text-to-image models. One approach is to include a broad range of fine-grained metrics, such as model bias (e.g., race, gender), aesthetics, reasoning capacity, toxicity, and multilingualism to name just a few [6, 13]. Nevertheless, two stand-out as the most commonly used: a) *quality*, also referred to as fidelity which measures the overall quality and realism of the generated images. and b) *alignment* which measures how well the generated image aligns with the input text prompt.

Current benchmarks [6, 11, 13, 18, 20, 24] focus mostly on enhancing the prompt set. Questions remain simple, binary answer questions. Even recent works [14] that collect more detailed feedback, e.g. artifact grounding, still rely on binary misalignment between text and image. This is problematic for several reasons:

Challenge during annotation. Simple, holistic questions about quality and alignment (as shown in Fig. 1, top) are manageable for short prompts. However, complex prompts present substantial difficulty for annotators, who often struggle with where to begin. As the task is highly subjective, this difficulty and ambiguity leads to inconsistent annotation practices across people and even within person across images, compromising annotation quality.

Presence-based alignment There is no clear definition on how alignment should be assessed. Binary assessment of alignment "Is there a dog in the image?" ignores the synthesis quality altogether. When judged independently, there is no difference between low quality synthetic images and actual photographs as long as concepts in the prompt are recognizable in the image. We ask ourselves: "What does it even mean that the alignment is perfect if the dog has three eyes or highly unrealistic fur?" This makes alignment by itself less useful and prevents using it as feedback signal for improving alignment in models.

Preventing deep analysis. With only two holistic labels for each image-text pair, it is challenging to identify the factors contributing to poor image quality or image-text misalignment in previous benchmarks hindering further deep dive on the model performance.

To address these issues, we propose *GenomeBench* a new framework for collecting human assessments of text-image pairs. We illustrate it in Fig. 1. Inspired by previous work in visual scene understanding and the concept of Scene Graphs, we parse prompts (Fig. 1(a)) into basic semantic components (Fig. 1(b)). A scene graph [12], a data structure describing scene contents, offers a straightforward encoding for image descriptions and a clear heuristic for assessing image quality and text-image alignment. By constructing node-specific questions (Fig. 1(c)), we significantly reduce the complexity for annotators. Departing from current practice, the node-specific questions do not ask for a binary rating but rather a gradual rating of alignment that implicitly embeds a notion of quality in the assessment.

We show that this approach has several advantages over existing practice: 1. *Higher inter-annotator agreement* due to transforming complex prompt inquiries into a sequential set of simpler questions, each focusing on individual semantic concepts. 2. *Higher correlation of alignment with overall image quality* than in existing automatic or manual metrics which mostly account for presence of semantic concept in the image. 3. *Explainability*, due to a scoring system that is fine-grained and that summarizes quality and alignment facilitating decision making (Fig. 1(d)) on model development.

Contributions are as follows: 1) we introduce a novel analysis framework for text-to-image generative models. This includes: a) a diverse prompt set with scene graph annotations (Sec. 3.1), providing fine-grain annotation on the quality of objects, attributes, relationship and numerals, b) a structured approach to guide human assessment (Sec. 3.2), and c) a transparent, explainable scoring system (Sec. 3.3). 2. we demonstrate the framework's advantages by applying it to recent text-to-image diffusion models (Sec. 4), revealing detailed insights into model performance, the relationship between image quality, text-image alignment, and complexity, among other aspects (Sec. 5).

2. Related Works

Reliable evaluation is crucial for the development and comparison of text-to-image (T2I) generative models. Various approaches have been proposed to assess these models,

focusing on different aspects of generative quality.

The two most commonly evaluated aspects are: a) *image quality*, also referred to as fidelity which measures the overall quality and realism of the generated images. and b) *alignment between the generated image and the input text* which assess how well the generated image aligns with the input text prompt.

While these two aspects are widely used, some researchers advocate for a broader range of fine-grained metrics, such as model bias (e.g., race, gender), aesthetics, reasoning capacity, toxicity, and multilingualism [6, 13].

In terms of human involvement, evaluation methods can be categorized as manual or automatic. *Human assessment methods* directly evaluate the generated images by humans. *Automatic assessment* rely on computational metrics which use pretrained neural networks to assess image quality and text-image alignment.

2.1. Human Assessment

Because existing automatic measures do not adequately reflect human judgments a standardized human evaluation protocol is proposed in [18]. [24] collect a human preference dataset by requesting users to rank multiple images and rate them according to their quality. [11] built a web application to collect human preferences by asking users to choose the better image from a pair of generated images. DrawBench, introduced in Imagen [20], employs human annotators for pairwise model comparisons, focusing on image fidelity and text-to-image alignment. PartiPrompts also targets MS-COCO's simplicity by offering prompts of varying complexities and explicitly labeling challenge levels. However, it has limitations in its single challenge label per prompt and simplistic evaluation questions.

Despite these valuable contributions, most existing works only use binary human ratings or preference ranking for construction of feedback/rewards, and lack the ability to provide detailed actionable feedback such as implausible regions of the image, misaligned regions, or misaligned keywords on the generated images. In response to this, RichHF-18K provides detailed human feedback on text-to-image generation [14]. Unfortunately annotations guidelines are unclear and the data is dominated by human faces.

2.2. Automatic Assessment

Perceptual metrics like Fréchet Inception Distance (FID) [8], CMMD [10] and Learned Perceptual Image Patch Similarity (LPIPS) [22] use pre-trained neural networks to assess the quality of generated imagery. However, these metrics rely on reference images and do not generalize well to evaluating the alignment between generated images and input text prompts.

To address this limitation, recent text-to-visual systems have predominantly reported using the CLIPScore [19], which measures the cosine similarity between the embedded image and text prompt. Yet, CLIP has been shown to struggle with reliably processing compositional text prompts [23, 26].

To mitigate this problem human-feedback approaches like ImageReward [24], PickScore [11] fine-tune vision-language models on large-scale human ratings collected for generated images. Divide-and-conquer approaches [17, 21] use large language models (LLMs) to decompose text prompts into simpler components for analysis. A notable technique within this framework is Question Generation and Answering (QG/A), exemplified by TIFA [9] and Davidsonian [7], where the text prompt is decomposed into QA pairs, and the alignment score is computed based on the accuracy of the answers generated by a VQA model. More reently, VQA [16] propose VQAScore, which utilizes a visual question-answering (VQA) model to determine if a generated image accurately depicts a given text prompt by answering a simple "Does this figure show 'text'?" question.

3. GenomeBench

GenomeBench consists of three main components: a scene graph annotated data corpus that we describe in Sec. 3.1, a mechanism to collect structured human assessments through guided question answering, described in Sec. 3.2 and a score that summarized image quality and alignment-to-text described in Sec. 3.3.

3.1. Data

GenomeBench prompts are sampled from two sources. The first set of prompts is coming from publicly available image-text datasets: MS-COCO [15], DrawBench [20] and PartiPrompts [25]. We complement these public sources with HIT, a corpus of prompts we have sourced internally, describing images of products sold online.

Inspired by the VisualGenome dataset [12], in these prompts we parse attributes, objects, relationships and numerals.

This is done in a stage-wise semi-automatic fashion. First, given a prompt, we pass it to an LLM[1] with instructions to return attributes, objects and actions. The result is then manually corrected by a human evaluator. This initial set of prompts tags is than enriched with additional fine grained subcategories of objects, attributes and relationships. We divide objects into: animate (humans, animals), products, context and abstract. Attributes are divided into color and material and relationships into associative, spatial, active, directional and qualitative. Finally, GenomeBench contains 301 unique prompts and 1809 unique tags[2]. In 1

[1] ChatGPT

[2] Object, attribute, relationship or numeral.

and 2 we show the data distribution by prompt source, tag source and tag-type. The difference between the first two comes from the higher tag density coming from HIT which generally contains more complex prompts compared to, for example, Parti. In Fig. 2 we show prompt complexity, measured by mean number of words and mean number of tags per source.

	Coco	Parti	HIT	Draw
Prompt	27.45	34.11	12.94	25.49
Tag	28.37	30.37	19.07	22.17

Table 1. Prompt and tag distribution by source in %.

Objects	Attributes	Relationships	Numerals
41.9	28.6	25.2	3.75

Table 2. Prompt and tag distribution by source in %.

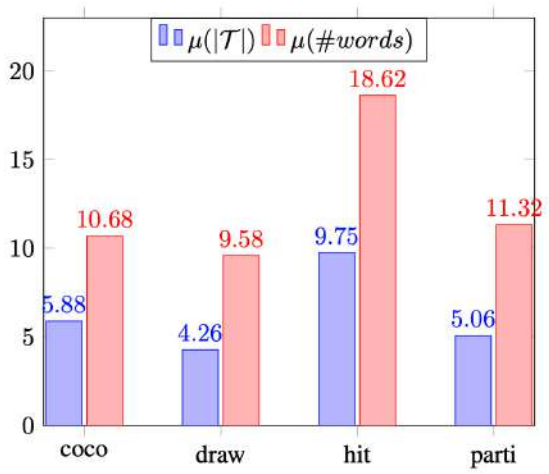

Figure 2. Prompt complexity, measured by mean number of tags and words, per source in GenomeBench.

3.2. Collecting Human Assessment

For facilitating human assessment collection we develop a set of new annotation tooling and task design.

3.2.1 Annotation Tooling

We developed a custom annotation tool for GenomeBench, which is based on Streamlit[3].

For a given prompt and its synthesized image, our interface starts with a holistic image appeal question and then traverses the prompt to rate each concept at a time, as shown in Fig. 3. The UI will dynamically change its question according the type of concepts (attributes, objects, relationships and numerals) being evaluated.

Tab. 3 shows the question template for each of the concepts. For each image-prompt pair, the image appeal question will be asked first, to get an unbiased holistic view irrespective of the text prompt. We choose to show the questions one at a time, since all questions at once likely lead to inconsistent attention shifts between questions among different annotators. All questions except for the Image Appeal question has "N/A - object missing" to cover the case when the object involved in the concept is missing. We also added "Cannot answer/Something is wrong" for each question in the case when the question does not make sense for the image.

3.2.2 Design of Annotation Task

The annotation task is designed as a randomized control experiment. Prompts, seeds and generation parameters are all fixed. The only variable is the generation model. This ensures fair comparison across all models.

Within each annotation batch, the annotator sees images from all models and all seeds across a set of prompts, with the display order randomized at image level. Such randomization ensures the internal standards shift across annotations is counterbalanced and uncorrelated with the results.

Each batch takes one to two hours to finish, with annotation results auto-saved. Annotators are allowed to work at their own pace without time limit.

3.3. Score

Let $\mathcal{D} = (X, Y)$ be a set of text and image pairs where X are natural language text strings and $Y \in \mathbb{R}^{W \times H \times C}$ are color images. We define a parser $\phi : X \rightarrow \mathcal{T}$ that assigns lexical tags $\mathcal{T}$ to the text x. The tags take values in a discrete set $\mathcal{T} \in \{\mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{N}\}$ of objects, attributes, relationships and numerals. Given the X, Y and $\mathcal{T}$ a set of scores are assigned for each tag. In practice, this is done by mapping the collected annotations (see Sec. 3.2) on a scale between 0 and 1. Finally, for each sample (x, y, T) drawn from $\mathcal{D}$ we define a general score:

$$S = \frac{\omega(y) + \alpha(x, y, \mathcal{T})}{2} \tag{1}$$

where $\omega(.)$ is an overall image quality score and $\alpha(.)$ is a text-to-image alignment and $S \rightarrow [0, 1]$. More specifically, α is defined by averaging over individual tags' alignment (object, attribute, relationships and numeral):

[3]https://streamlit.io/

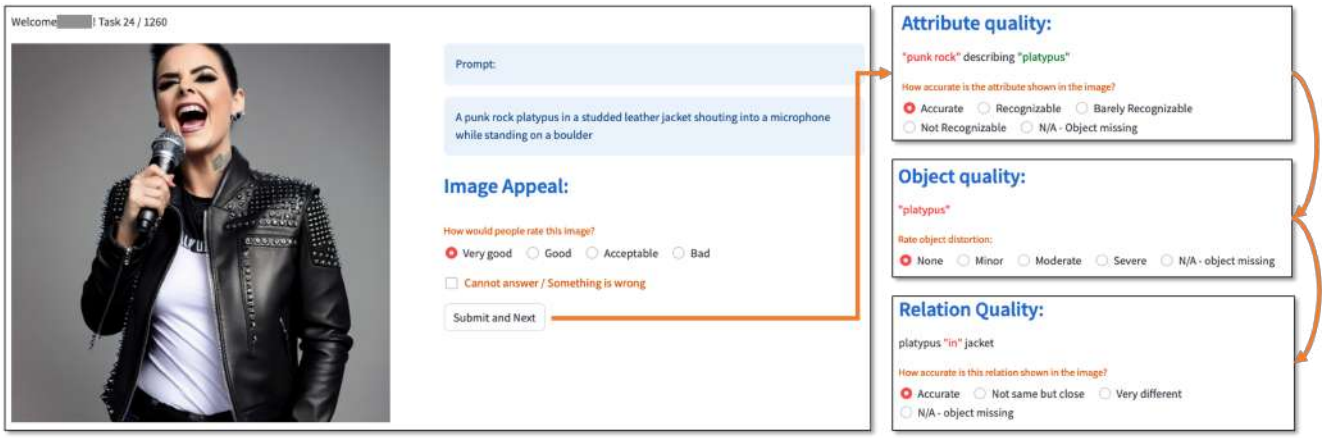

Figure 3. Annotation UI for GenomeBench. The example prompt used here is "A punk rock platypus in a studded leather jacket shouting into a microphone while standing on a boudler". Annotators will see the synthetic image on the left and questions on the right. Each question evaluates the quality of a concept.

Concept	Question Template	Response
image appeal	How would people rate this image?	Very good, Good, Acceptable, Bad
object	<obj> - Rate object distortion	None, Minor, Moderate, Severe
attribute	<attr> describing <obj> - How accurate is the attribute shown in the image?	Accurate, Recognizable, Barely Recognizable, Not Recognizable
relationship	<obj1> <rel> <obj2> - How accurate is this relation shown in the image?	Accurate, Not same but close, Very different
numeral	<num> describing <obj> - Is this number accurately shown?	Exactly, A bit less, A bit more, A lot less, A lot more

Table 3. Questions and response for each concept.

$$\alpha(x,y,T) = \frac{1}{|\mathcal{T}|}\Big(\sum_{o\in\mathcal{O}}\alpha(o,x) + \sum_{a\in\mathcal{A}}\alpha(a,o,x) + \sum_{r\in\mathcal{R}}\alpha(r,o_1,o_2,x) + \sum_{n\in\mathcal{N}}\alpha(n,o,x)\Big) \quad (2)$$

4. Experiments

To demonstrate its effectiveness we collect human annotations over images generated with the GenomeBench prompts. The sample is random and it preserves the distribution characteristics described in 3.1. Seven human annotators, 6 men, 1 woman, ages from 30 to 50, university educated, from diverse ethnic backgrounds, living in different locations in North America and Europe, were instructed to complete the annotations tasks as described in 3.1.

We include several publicly available text-to-image diffusion models in our analysis: Realistic Vision [1], Dreamlike [2], Deliberate [3], Stable Diffusion 2.1 [4] and OpenJourney [5].

We synthesize images for each prompt model combination. We use a standard inference pipeline[4] with a fixed setup. We synthesize the same three seeds per combination which results in 4500 unique images. We detail and comment the results of this study in the next section.

5. Results and Analysis

In this section we discuss overall results. We compare our proposal with baseline human assessment in Sec. 5.1 and with automatic assessment in Sec. 5.2. We then demonstrate deep analysis on a set of recent text-to-image generative models in Sec. 5.3. Finally in Sec. 5.4 we study the relationship between quality, alignment and prompt complexity and we conclude with examples in Sec. 5.5.

5.1. Comparison with Baseline Human Assessment

We perform a consistency analysis between the proposed alignment (α) and the baseline human assessment. For building a baseline we have asked two annotators of the initial benchmark to annotate image-to-text alignment. This corresponds to the baseline scenario (no guided QA; as in

[4]Available on https://huggingface.co/

DrawBench and PartiPrompts). We aggregate assessments of this baseline into model rankings per annotator and we compute consistency for model ranking among annotators. Similarly we compute mean and standard deviation consistency among all pairs of annotators from GenomeBench. In Table 4 we show Kendall rank correlation coefficient of model ranking. Notice how structuring human assessment achieves higher agreement on model ranking.

Methods	α
Baseline	0.06
GenomeBench	0.2 $\pm$0.08

Table 4. Inter-annotator agreement measured by Kendall tau.

5.2. Comparison with Automatic Assessment

In Table 5 we show correlation between baseline human assessment of quality and α, VQA [16], DSG [7] and TIFA [9] (three automatic alignment assessment metrics). Results clearly show that automatic alignment assessment correlates poorly with image quality. In other words, current state-of-the-art in automatic text-to-image alignment might account for presence in the image, i.e. "Is there a dog in the image?" but ignores quality altogether. By specifically asking annotators about the synthesis quality of semantically meaningful items in the text, our score α not only detects if a semantic concept is present in the image but also measure its rendition quality.

5.3. Dissecting Diffusion Models' Performance

We proceed by demonstrating the deeper analysis a structured score can achieve. We show in Table 6 the proposed score and the models studied and its breakdown into the quality ω and alignment α components. We showing in Table 7 quality alignment based on object category, in Table 8 we analyse alignment based on attribute category while and Table 9 we analyse alignment based on relationship category.

The benchmark proposed is a method of assessing behaviour but causes for such behaviour are in the model development itself: data, architecture, training protocol, model capacity etc. It is difficult to speculate on why a certain model is better than another. Nevertheless, the proposed benchmark helps the interested practitioner to create hypothesis and guides toward possible action items. This comes in contrast to the current holistic assessment that will only conclude that "model X is better than model Y" without offering hints on how could model Y be improved.

We will use our study as an example. From Tables 6,7, 8,9 one can make a set of interesting observations about the models studied. First, animate objects (humans, animals) have the lowest quality among objects. It is an intuitive result and it confirms the anecdotal experience of practitioners observing extra-fingers or legs. This is for sure due to the

Metric	VQA	DSG	TIFA	Ours
Correlation	0.158	0.133	0.075	0.417

Table 5. Correlation with holistic human perception of quality.

Model	Quality	Alignment				Score
		Obj	Attr	Rel	Num	
RV	**0.70**	**0.65**	**0.78**	**0.81**	0.73	**0.72**
Del	0.67	0.62	**0.78**	0.76	0.70	0.71
Drm	0.64	0.61	**0.78**	0.78	0.47	0.65
SD 2.1	0.55	0.58	0.75	0.77	**0.79**	0.64
OJ	0.53	0.57	0.74	0.74	0.73	0.61

Table 6. Overall score and its quality and alignment components across models. Best score in **bold**. RV: Realistic Vision, Del: Deliberate, Drm: Dreamlike, SD: Stable Diffusion, OJ: OpenJourney

Model	Animate	Context	Products
RealisticVision	0.55	**0.81**	**0.68**
Deliberate	0.53	**0.81**	0.65
Dreamlike	**0.58**	0.78	0.59
StableDiffusion 2.1	0.50	0.70	0.62
OpenJourney	0.45	0.77	0.62
Theia 1.0	0.40	0.68	0.63

Table 7. Model performance over object type alignment. Best score in **bold**.

Model	Color	Material
RealisticVision	**0.80**	0.81
Deliberate	0.79	**0.82**
Dreamlike	0.77	0.75
StableDiffusion 2.1	0.70	0.79
OpenJourney	0.77	0.59
Theia 1.0	0.68	[0.84]

Table 8. Model performance over attribute type alignment. Best score in **bold**.

Model	Spatial	Qualitative	Associative
RV	**0.56**	**0.90**	0.74
SD 2.1	0.51	0.68	**0.81**
Dreamlike	0.53	0.88	0.70
Deliberate	0.55	0.67	0.68
OpenJourney	0.51	0.70	0.72
Theia 1.0	0.58	0.97	0.73

Table 9. Model performance over relationship type alignment. Best score in **bold**. RV: Realistic Vision, SD: Stable Diffusion

higher variation in appearance of articulated, non-rigid objects compared to the lower degree of variation of a rigid objects. Practitioners should invest considerably more in curating large and diverse images of humans and animals. Additionally, observers are highly sensitive to any structural inconsistency of the human body and face. Artifacts there would heavily weight down any general perception of quality. Second, it seems that context is the easiest type of object to model. This of course, might also mean that humans are far less sensitive to the quality of the background. Any good model should first focus on getting the salient foreground objects right. Third, spatial relationships are difficult to grasp by these models. They represent abstract conceptual knowledge that the model has to learn. This is the classic example of the "astronaut on a horse". From what we have observed spatial relationships are particularly tricky between animate objects, e.g. "a cat on a dog". Finally we add an observation not transparent from the results. Counting is not particularly bad in the benchmark because low numbers (two-three) are over-represented. Models like these fail almost always with higher numbers. This is a clear indication of learning by association and suggests that the textual embedding is less than ideal.

5.4. On the Relationship between Quality, Alignment and Prompt Complexity

We now turn to the relationship between quality and alignment. For this purpose, in Tab. 10 we show the correlation of the various type of alignment: overall alignment α, object-alignment $\alpha_{\mathcal{O}}$, attribute-alignment $\alpha_{\mathcal{A}}$, relationship-alignment $\alpha_{\mathcal{R}}$ and quality ω. Notice how, among the alignment components, the most correlated with quality is by far object-alignment. This is intuitive and suggests that when human judge image quality, they mainly focus on object quality.

$corr(\omega, \alpha)$	$corr(\omega, \alpha_{\mathcal{O}})$	$corr(\omega, \alpha_{\mathcal{A}})$	$corr(\omega, \alpha_{\mathcal{R}})$
0.51	0.71	0.29	0.18

Table 10. Correlation between quality and the different components of alignment.

Next, we look into prompt complexity. Since the recent emergence of potent text-to-image models there was frantic activity in the community for searching for prompts that create best images. More often than not these prompts would be very detailed and quite intricate. We have now, for the first time a clear measure of both image quality and prompt complexity. In Table 11 we show that in fact there is almost no correlation between prompt complexity as measured by the number of tags $|\mathcal{T}|$ and α and ω across all models studied. This measurement is model independent. This is in a sense surprising as one would expect that a model finds it easier to synthesize good quality images from simpler prompts. Similarly, it seems alignment does not necessarily decrease with more complex prompts. When taken per model, it might be that specific models express stronger dependencies than others.

$corr(\|\mathcal{T}\|, \alpha)$	$corr(\|\mathcal{T}\|, \omega)$
0.10	0.01

Table 11. Correlation between prompt complexity, i.e. mean number of tags, and quality and alignment per prompt.

5.5. Qualitative Results

To further illustrate the proposed framework in Fig. 4 we show a sample from each model for four different prompts. For each particular sample, we specify the quality ω and alignment α as they were obtained during our study from actual human assessments. If we are to focus on the first prompt, the quality is pretty much the same, with the notable exception of sample (f) which is rated higher. If on the other hand we look at the alignment scores, there are great differences. On one side, (a) and (e) are poorly aligned. Notice how in both of the "a man" is entirely missing. At the either side of the spectrum, (b) comes closest to perfect alignment: not only we have all the object present but in this case also the relationship "standing on (man, cart)" is relatively aligned. Consider now the second example. Image (i) is considered to be the most appealing by the raters. But it is by no means the most aligned to the prompt. Even if none of the images is perfectly aligned the one the is the most aligned is image (h) which happens to be the only one that contains both "cat" and "dogs". The obvious source of misalignment of course comes from the numerals. In the third example the prompt is particularly challenging. This immediately obvious from the low alignment scores across models. On the quality, image (l) scores very high. it is a nice, high quality, high contrast image. nevertheless, even in this case the alignment to the prompt is minimal. Finally, in the last set of examples we get considerable variance along both dimensions. By far the worst sample (s) for obvious reasons. Example (p) scores highest overall, both on quality and alignment. You might wonder why are (r) and (t) not scored higher. At a closer look one might notice that what you see is the actual Earth (as judged by the details of the planetary surface) and the moon is not even present. The rest of course is captured in the quality of the object "International Space Station".

5.6. Limitations

Despite the significant contributions and insights gained from this study, it is essential to acknowledge several inherent limitations, some unique to GenomeBench, some shared among the human-in-the-loop benchmarks.

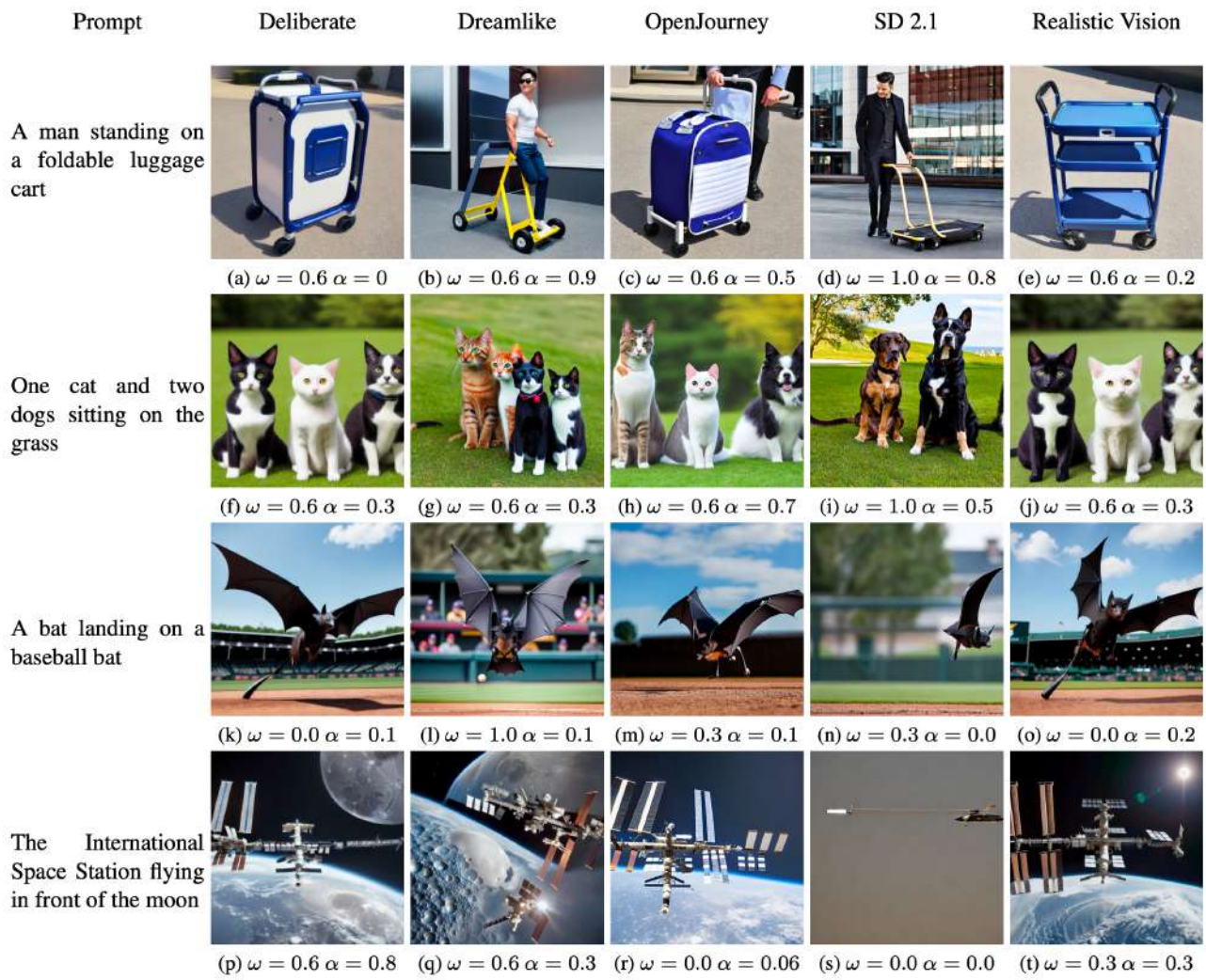

Figure 4. Examples of synthesized images and their associated quality and alignment scores.

Most importantly, the GenomeBench score requires considerable human question answering. Although the questions are now more well defined and analysis more powerful, annotators still need to spend more time on each image-text pairs, especially when the prompt is complex. Such challenge is also shared with previous human assessment as annotators need to carefully consider each image without explicit guidance.

Second, the need of human annotation also means that given a newly developed model it is not immediate to calculate the model performance and conduct model comparisons as it is the case with automatic metrics assessment. This shortcoming is shared among all the benchmark systems that requires manual annotation.

Third, if new prompts are requested, additional tag parsing is needed before the use of GenomeBench. This can be mitigated by appropriate planing on prompt set expansion.

6. Conclusion

In this paper we proposed GenomeBench, a comprehensive framework that includes a prompt dataset with semantically rich annotations and a system to structure human assessment through detailed questioning over text-image quality and alignment. This showed superior inter-annotator agreement over baseline binary human assessment and higher correlation between alignment and quality when compared to automatic assessment. Furthermore, the implicit fine-grained structure, allows for in-depth analysis of generative models, shedding light on their ability to generate realistic and high-quality samples.

References

[1] https://huggingface.co/SG161222/Realistic_Vision_V1.4/tree/main. Accessed:

2023-11-15. 5

[2] https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0. Accessed: 2023-11-15. 5

[3] https://huggingface.co/XpucT/Deliberate. Accessed: 2023-11-15. 5

[4] https://huggingface.co/stabilityai/stable-diffusion-2-1. Accessed: 2023-11-15. 5

[5] https://docs.midjourney.com/docs/model-versions. Accessed: 2023-11-15. 5

[6] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models, 2023. 2, 3

[7] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 3, 6

[8] 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. *Advances in neural information processing systems*, 30, 2017. 3

[9] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 3, 6

[10] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 3

[11] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023. 2, 3

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2, 3

[13] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy S Liang. Holistic evaluation of text-to-image models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 69981–70011. Curran Associates, Inc., 2023. 2, 3

[14] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[15] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[16] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 3, 6

[17] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023. 2, 3

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[20] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[21] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36:70799–70811, 2023. 3

[22] Jake Snell, Karl Ridgeway, Renjie Liao, Brett D Roads, Michael C Mozer, and Richard S Zemel. Learning to generate images with perceptual similarity metrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4277–4281. IEEE, 2017. 3

[23] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 3

[24] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagere-

ward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 2, 3

[25] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2, 3

[26] 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 *The Eleventh International Conference on Learning Representations*, 2023. 3