Comparing Text-to-Image Consistency between Abstract and Realistic Objects

Marc Hollinger

Abstract

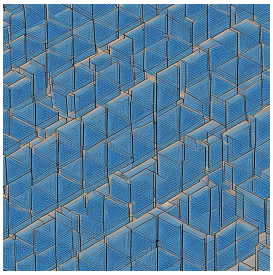
Text-to-image generation is non-deterministic and different wordings of a prompt can lead to different outcomes. However, if the prompt requires an object to be present in the scene, it should be present regardless of the exact wording of the prompt. In other words, images generated from different descriptions of the same scene should be similar. This can be quantified by creating semantically equivalent variations of a prompt, scoring the resulting images on faithfulness and analyzing the variation in scores within a semantically equivalent group. Using this variation as a consistency score, it can be shown that across different versions of Stable Diffusion, the consistency is worse for prompts describing scenes with abstract geometric shapes than it is for prompts describing realistic scenes with complex objects.

1 Introduction

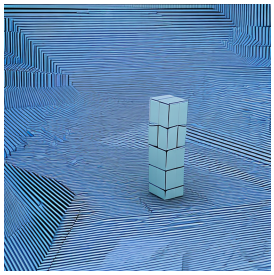
In many image generation applications, prompts request one or multiple objects to be present in the image (Metzer et al., 2022; Kumari et al., 2023; Höllein et al., 2024). Regardless of the exact formulation, the model should generate a picture where the requested object and background are present. Increasing the likelihood that the image is faithful could be useful for both 2D and 3D applications where the text-to-image model needs to produce an image of an object (Metzer et al., 2022; Poole et al., 2022).

An observation I made with Stable Diffusion (SD) is that when trying to generate images of abstract objects like a pyramid or a cube, the object is sometimes completely unrecognizable while this rarely happened for more complex objects like a butterfly or a car. Compare the results in Figure 2 with Figure 1.

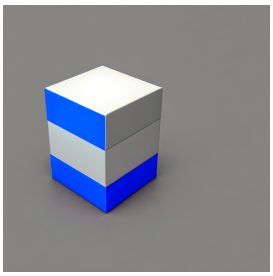
This benchmark seeks to quantify the perceived difference in consistency between abstract and realistic objects.



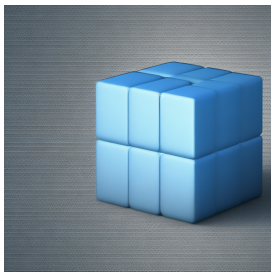
(a) "A blue 3d-rendered cube."



(b) "A 3D-rendered cube in blue."



(c) "A cube rendered in 3D with a blue color."



(d) "A blue cube created with 3D rendering."

Figure 1: Each prompt variation requests the same scene but the generated images vary widely in faithfulness. Images generated by SD2 (num_inference_steps=50; guidance_scale=7.5). (a) does not depict an object at all. An object is present in (b) but it is not a cube. The object in (c) and (d) are cuboids but only (d) looks like all side lengths are approximately equal. The requested color consistently appears in the images but there is often some kind of texture to the object that was not mentioned in the prompt.

2 Method

Scoring the quality of artificially generated images is usually done on two main criteria: image quality and text-image alignment (Zhang et al., 2023). For this benchmark, the focus is on text-image alignment rather than image quality. The goal is to measure if the requested object is present in the image with less importance placed on how accurate the object itself looks. To measure the alignment between image and prompt, I used ALIGNscore as described in Saxon et al. (2024) based on Jia

(a) "A butterfly resting on a flower."

(b) "A butterfly sitting atop a flower."

(c) "A butterfly alighting on a flower."

(d) "A butterfly settled on a bloom."

Figure 2: Realistic object generated by SD2 (num_inference_steps=50, guidance_scale=7.5). The requested object and background are present in all output images regardless of quality.

et al. (2021) but any metric that produces a numeric score can be used, as the benchmark is based on variability rather than absolute scores.

I used GPT-4o (temperature=0.7; June 2024 version) to generate 40 prompts requesting real world objects as well as 40 prompts requesting simplified 3d-rendered objects. For simplicity's sake most prompts describe a scene only containing a single object.

To test consistency across different formulations of a prompt, I used GPT-4o to reformulate each prompt in five semantically equivalent ways. The variations were eye-checked to ensure that the content is approximately equivalent.

The procedure to evaluate a model is as follows: For every object o_i, the model uses prompt variation L_{ij} to generate image I_{ij}, where $i \in \{1, 2, \ldots, n\}$, $j \in \{1, 2, \ldots, m\}$. n and m are the number of objects and the number of variations per object respectively. In this benchmark the first n_a objects belong to the abstract category and the next n_r belong to the realistic category, where $n_\text{a} = n_\text{r} = 40$ and $n_\text{a} + n_\text{r} = n$.

The alignment between the generated images and the prompts is calculated using a chosen faithfulness metric, resulting in five scores $v_{i,j}$ for each object o_i.

$$v_{i,j} = \text{ALIGNScore}\left(I_{ij}, L_{ij}\right) \tag{1}$$

To get a score s_i for object o_i a metric is used to capture the variability of scores belonging to the same object. By default the metric chosen is the sample standard deviation.

$$s_i = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(v_{ij} - \overline{v_i})} \tag{2}$$

A system with high consistency should show a lower variance in alignment scores across the five semantically equivalent prompts. High variance indicates that the model is not consistently able to produce a faithful output for the object requested in the prompt.

To compare the performance on abstract objects to the performance on photorealistic objects, a score for each category is calculated by taking the mean of the object scores belonging to that category.

$$S_\text{a} = \frac{1}{n_\text{a}}\sum_{i=1}^{n_\text{a}} s_i \tag{3}$$

$$S_\text{r} = \frac{1}{n_\text{r}}\sum_{i=n_\text{a}+1}^{n} s_i \tag{4}$$

where $n = n_\text{a} + n_\text{r}$ and, for this benchmark, $n_\text{a} = n_\text{r} = 40$.

A good benchmark result means that the mean score is similar for both realistic and simplified prompts. The benchmark score is the difference between the means of the two categories.

$$S_{final} = S_r - S_a \tag{5}$$

A positive score means that the model shows greater variability when generating realistic images while a negative score means the model shows greater variability when generating images of abstract objects. Because a deviation in either direction is undesirable, a lower absolute benchmark score is considered better.

3 Results

Using the benchmark on three different versions of StableDiffusion (2.0, 2.1, 3-medium) shows that there does seem to be a difference between simplified and realistic objects. The mean consistency

score (mean of std score) is higher for abstract objects than for realistic objects for every version of SD tested. Moreover, the variation of the consistency score is higher for abstract objects than for realistic objects. Overall, the mean consistency on simplified objects increases with later versions.

Comparing with DALL-E 3 reveals that it performs very similarly to SD3-medium but the mean variability is higher for DALL-E 3 in both categories. (Fig. 3). Comparing only based on the difference between categories shows that DALL-E 3 performs better on the standard benchmark as well as when using the median as metric aggregation metric (Table 1). When using the minimum to aggregate the five object scores SD3-medium performs better.

model-id	**metric**		
	std	min	med
SD3-medium	-0.86	**-2.10**	-3.19
DALL-E 3	**-0.54**	-2.20	**-3.12**

Table 1: Comparing SD3-medium with DALL-E-3 on benchmark score with different metrics used in equation 2. Since this score is the difference between categories, a low absolute score is desired. The most pronounced difference in scores occurs when using the standard deviation. This indicates that the difference is because of variability rather than direct faithfulness.

4 Conclusion

The results across different versions of SD show that the variability difference between abstract and realistic objects decreases as the models become more performant in general tasks. SD3-medium manages to outperform DALL-E 3, which was much more consistent than previous SD versions especially on abstract objects.

It seems likely that this discrepancy will disappear with more advanced models. However, something that even the most consistent model produces is texture or details that were not requested. This mainly happens on abstract objects and is likely caused by training data that shows similar shapes but with texture and details.

An interesting task for future research could be to find a mechanism by which the model is able to produce an image "bottom-up" only adding details if explicitly requested.

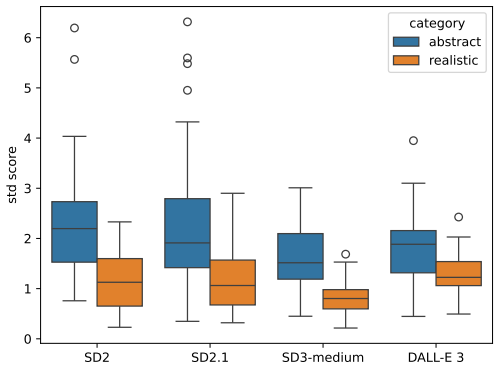


Figure 3: Distribution of the standard deviations by model and task type. All versions of SD used the same hyperparameters (num_inference_steps=50, guidance_scale=7.5) The variability is consistently higher for abstract objects than it is for realistic objects. The mean decreases with more recent models, showing that variability decreases with increasing model performance. The variability of scores across all objects within a category shows a similar trend but increases slightly between versions 2 and 2.1. For DALL-E 3 (size="1024x1024"; style="vivid"; quality="standard"), the difference between categories is very similar to SD3-medium while the means of both categories are higher for DALL-E 3 indicating higher variability.

References

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. 2024. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5043–5052.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2022. Latent-nerf for shape-guided generation of 3d shapes and textures.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion.

Michael Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. 2024. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (ts2).

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023. Text-to-image diffusion models in generative ai: A survey.