Getting it *Right*: Improving Spatial Consistency in Text-to-Image Models : Supplementary Material

Agneet Chatterjee [*1], Gabriela Ben Melech Stan [*2], Estelle Aflalo[2], Sayak Paul[3], Dhruba Ghosh[4], Tejas Gokhale[5], Ludwig Schmidt[4], Hannaneh Hajishirzi[4], Vasudev Lal[2], Chitta Baral[1], and Yezhou Yang[1]

[1] Arizona State University
[2] Intel Labs
[3] Hugging Face
[4] University of Washington
[5] University of Maryland, Baltimore County

In this supplementary material, we present additional quantitave and qualitative results from our dataset and method. We discuss fine-grained FaithScore evaluations of the SPRIGHT captions, along with ways to improve the caption quality and its impact on models that support longer token limits. We present the GPT-4 (V) prompt used for evaluation and discuss the limitations of our current work. Lastly, we cover the contributions of each author in this work.

1 Results on T2I-CompBench

As shown in Table 1, we achieve state of the art performance on the spatial score in the widely accepted T2I-CompBench benchmark. The significance of training on images containing a large number of objects is emphasized by the enhanced performance of our models across various dimensions in T2I-CompBench. Specifically, we enhance attribute binding parameters such as *color* and *texture*, alongside maintaining competitive performance in *non-spatial* aspects.

2 FaithScore Evaluations

Table 2 presents the detailed breakdown of the FaithScore evaluations conducted on the SPRIGHT captions, with the spatially-focused relationships being 83.6% correct, on average.

3 CLIP Token Limit

The longer SPRIGHT captions better utilize the CLIP 77-token limit; ground truth and SPRIGHT captions have an average of `14.95` and `81.43` tokens, respectively. Furthermore, T2I models with longer context lengths and multiple

* Equal contribution. Correspondence to agneet@asu.edu

Table 1: Results on the T2ICompBench Benchmark. a) We achieve state of the art spatial score, across all methods, by efficient fine-tuning on only 444 images. b) Despite not explicitly optimizing for them, we find substantial improvement and competitive performance on attribute binding and non-spatial aspects.

Method	Attribute Binding			Object Relationship	
	Color	Shape	Texture	Spatial	Non-Spatial
SD 1.4	0.3765	0.3576	0.4156	0.1246	0.3079
SD 2	0.5065	0.4221	0.4922	0.1342	0.3096
Composable v2	0.4063	0.3299	0.3645	0.0800	0.2980
Structured v2	0.4990	0.4218	0.4900	0.1386	0.3111
Attn-Exct v2	0.6400	0.4517	0.5963	0.1455	0.3109
GORS	0.6603	0.4785	0.6287	0.1815	0.3193
DALLE-2	0.5750	0.5464	0.6374	0.1283	0.3043
SDXL	0.6369	0.5408	0.5637	0.2032	0.3110
PixArt-Alpha	0.6886	0.5582	0.7044	0.2082	0.3179
Kandisnky v2.2	0.5768	0.4999	0.5760	0.1912	0.3132
DALL-E 3	0.8110	0.6750	0.8070	-	-
Ours (<500 images)	0.6251	0.4648	0.5920	**0.2133**	0.3132

Table 2: FAITHScore caption evaluation of our SPRIGHT dataset. On a sample of 40,000 captions, SPRIGHT obtains an 88.9% accuracy, comparable with the reported 86% and 94% on LLaVA-1k and MSCOCO-Captions, respectively. On the subset of atomic claims about spatial relations, SPRIGHT is correct 83.6% of the time.

Category	# Examples	Accuracy (%)
Overall FAITHScore	—	88.9
Entities	149,393	91.4
Relations	167,786	85.8
Colors	10,386	83.1
Counting	59,118	94.5
Other	29,661	89.0
Spatial	45,663	83.6

text encoders such as PixArt-Sigma and SD3 can take full advantage of our captions and training technique: we fine-tune PixArt-Sigma (token limit = `300`) on SPRIGHT and obtain a spatial score of `0.2501`.

4 Improvements in Captioning

While our work is to explore the impact of spatially focused captions, we find that improvements in caption quality can be achieved through stronger models like LLaVA-1.6-34B, GPT-4(V) or GPT-4o. To validate this, we conduct a human

study (n=3) on 100 CC-12M images, comparing re-captioning performance of LLaVA-1.5-13B and LLaVA-1.6-34B, and find an improvement from 63% to 78%.

5 System Prompt for GPT-4 Evaluation

```
You are part of a team of bots that evaluates images and their
captions. Your job is to come up with a rating between 1 to 10
to evaluate the provided caption for the provided image. Consider
the correctness of spatial relationships captured in the provided
image. Return the response formatted as a dictionary with two
keys: 'rating', denoting the numeric rating, and 'explanation',
denoting a brief justification for the rating.

The captions you are judging are designed to stress-test image
captioning programs, and may include:
1. Spatial phrases like above, below, left, right, front, behind,
background, foreground (focus most on the correctness of these
words).
2. Relative sizes between objects such as small & large, big &
tiny (focus on the correctness of these words).
3. Scrambled or misspelled words (the image generator should
produce an image associated with the probable meaning). Make a
decision as to whether or not the caption is correct, given the
image.

A few rules:
1. It is ok if the caption does not explicitly mention each object
in the image; as long as the caption is correct in its entirety,
it is fine.
2. It is also ok if some captions don't have spatial
relationships; judge them based on their correctness. A caption
not containing spatial relationships should not be penalized.
3. You will think out loud about your eventual conclusion. Don't
include your reasoning in the final output.
4. Return the response formatted as a Python-formatted dictionary
having two keys: 'rating', denoting the numeric rating, and
'explanation', denoting a brief justification for the rating.
```

6 Comparing COCO-30K and Generated Images

In Figure 1, we compare images from COCO, baseline Stable Diffusion and our model. We find that the generated images from our model adhere to the input prompts better, are more photo-realistic in comparison to the baseline.

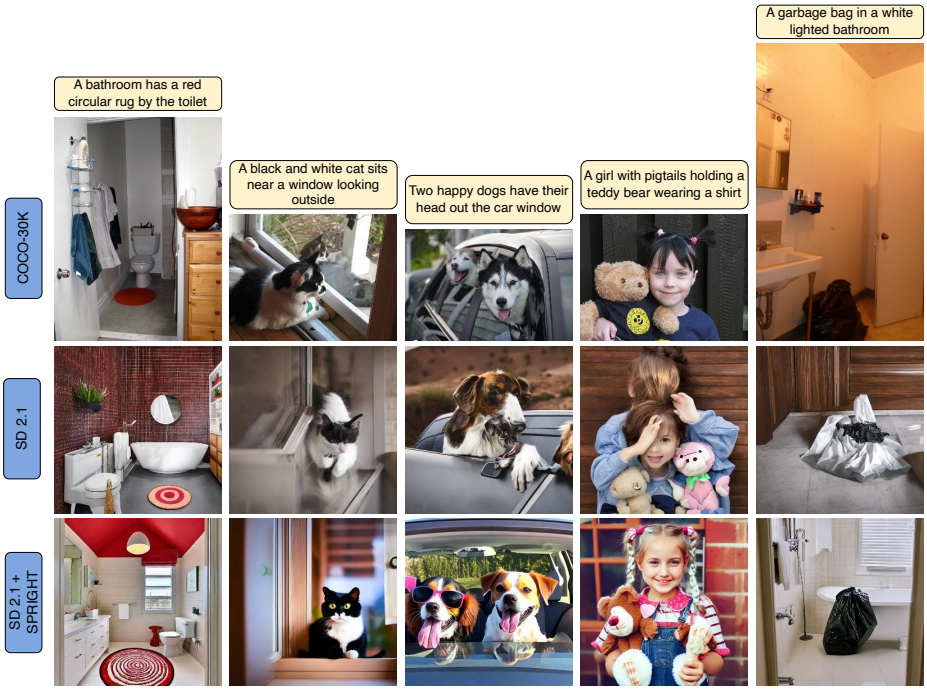

Fig. 1: Illustrative examples comparing ground-truth images from COCO and generated images from Baseline SD 2.1 and our model. The images generated by our model exhibit greater fidelity to the input prompts, while also achieving a higher level of photorealism.

7 Additional Examples from SPRIGHT

Figure 2 and 3 demonstrate a few correct and incorrect examples present in SPRIGHT. While most relationships are accurately described in the captions, on some instances the model struggles to capture the precise spatial nuance.

8 Additional Illustrations

Figure 4 shows images generated by our model based on prompts from T2I-CompBench, whereas Figure 5 demonstrates that for a given prompt, our model consistently produces spatially accurate images. Figure 6 presents example images generated from the VISOR benchmark.

9 Limitations

Since SPRIGHT is a derived dataset, it inherits the limitations of the original datasets. We refer the readers to the respective papers that introduced the original datasets for more details. As shown in our analysis, the generated synthetic

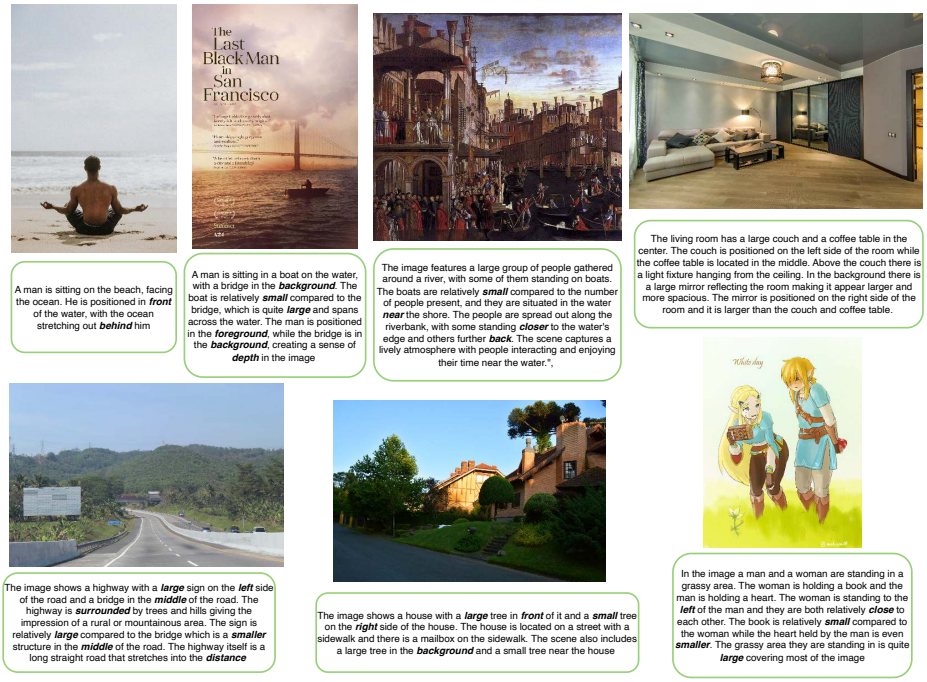

Fig. 2: Illustrative examples from the SPRIGHT dataset, where the captions are correct in its entirety; both in capturing the spatial relationships and overall description of the image. The images are taken from CC-12M and Segment Anything.

captions are not a 100% accurate and could be improved. The improvements can be achieved through better prompting techniques, larger models or by developing methods that better capture low-level image-text grounding. However, the purpose of our work is *not* to develop the perfect dataset, it is to show the impact of creating such a dataset and its downstream impact in improving vision-language tasks. Since our models are a fine-tuned version of Stable Diffusion, they may also inherit their limitations in terms of biases, inability to generate text in images, errors in generating correct shadow patterns. We present our image fidelity metrics reporting FID on COCO-30K. COCO-30K is not the best dataset to compare against our images, since the average image resolutions in COCO is lesser than those generated by our model which are of dimension 768. Similarly, FID largely varies on image dimensions and has poor sample complexity; hence we also report numbers on the CMMD metric.

10 Author Contributions

AC defined the scope of the project, performed the initial hypothesis experiments and conducted the evaluations. GBMS led all the experimental work and customized the training code. EA generated the dataset, performed the dataset

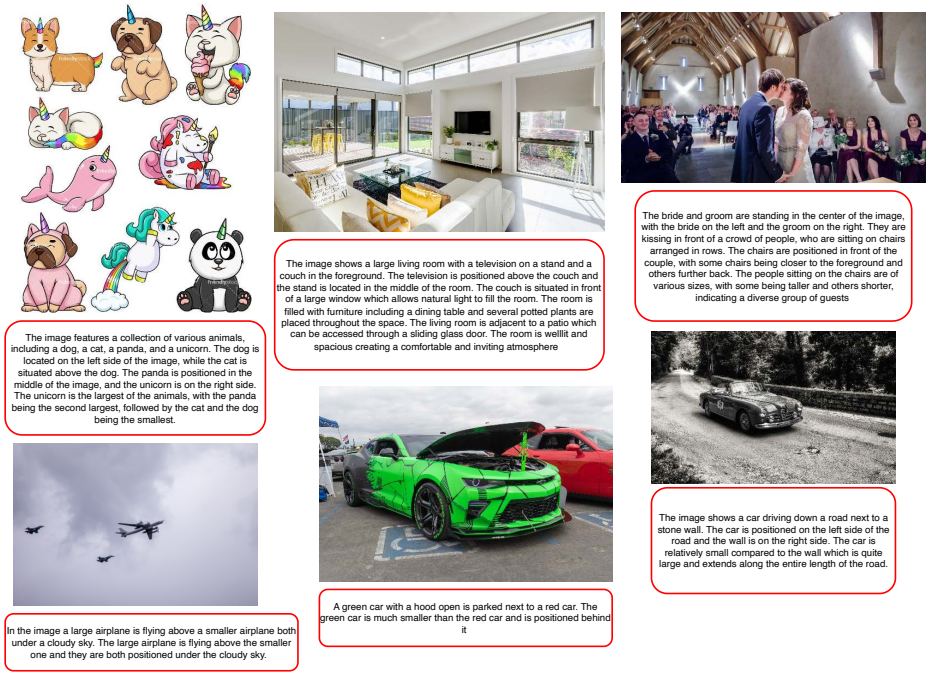

Fig. 3: Illustrative examples from the SPRIGHT dataset, where the captions are not completely correct. The images are taken from CC-12M and Segment Anything.

and relevancy map analyses. SP took part in the initial experiments, suggested the idea of re-captioning and performed few of the evaluations and analyses. DG suggested the idea of training with object thresholds and conducted the FAITH-Score and GenEval evaluations. TG initiated the discussions on spatial failures of T2I models and provided consultation on experiments. VL, CB, and YZ co-advised the project, initiated and facilitated discussions, and helped shape the the goal of the project. AC and SP wrote the manuscript in consultation with TG, LW, HH, VL, CB, and YZ. All authors discussed the result and provided feedback for the manuscript.

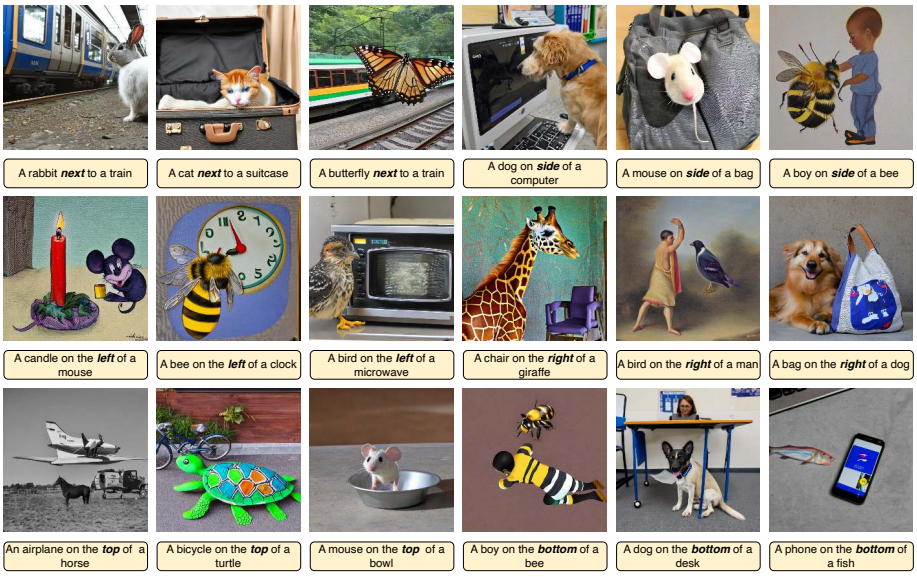

Fig. 4: Illustrative examples from our model, as described in Section 4.1, on evaluation prompts from the T2I-CompBench benchmark.

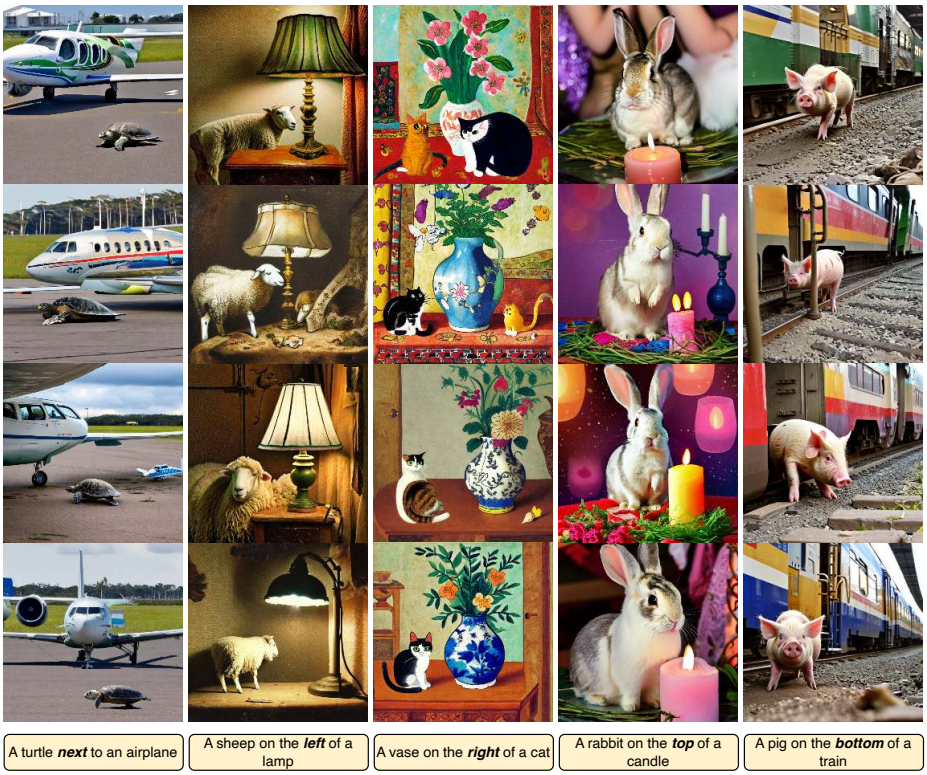

Fig. 5: Generated images from our model, as described in Section 4.2, on evaluation prompts from T2I-CompBench. We find that for a given text prompt, our model consistently generates spatially accurate images.

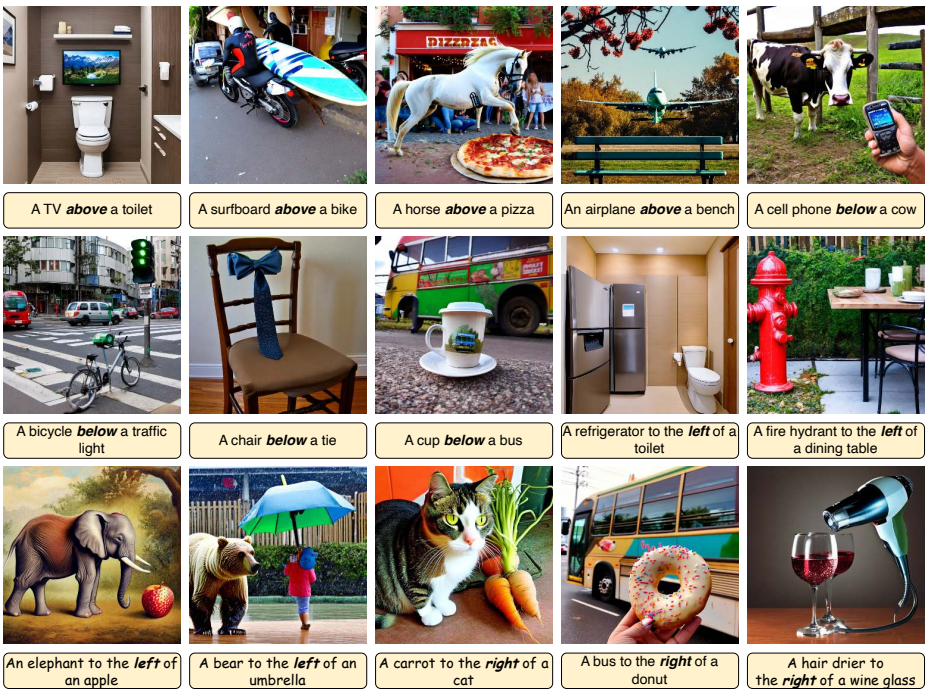

Fig. 6: Generated images from our model, as described in Section 4.2, on evaluation prompts from the VISOR benchmark.