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

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

Abstract. One of the key shortcomings in current text-to-image (T2I) models is their inability to consistently generate images which faithfully follow the spatial relationships specified in the text prompt. In this paper, we offer a comprehensive investigation of this limitation, while also developing datasets and methods that support algorithmic solutions to improve spatial reasoning in T2I models. We find that spatial relationships are under-represented in the image descriptions found in current vision-language datasets. To alleviate this data bottleneck, we create SPRIGHT, the first spatially focused, large-scale dataset, by re-captioning 6 million images from 4 widely used vision datasets and through a 3-fold evaluation and analysis pipeline, show that SPRIGHT improves the proportion of spatial relationships in existing datasets. We show the efficacy of SPRIGHT data by showing that using only ∼0.25% of SPRIGHT results in a 22% improvement in generating spatially accurate images while also improving FID and CMMD scores. We also find that training on images containing a larger number of objects leads to substantial improvements in spatial consistency, including state-of-the-art results on T2I-CompBench with a spatial score of 0.2133, by fine-tuning on *<500* images. Through a set of controlled experiments and ablations, we document additional findings that could support future work that seeks to understand factors that affect spatial consistency in text-to-image models. Project page : `https://spright-t2i.github.io/`

Keywords: Text to Image Generation · Spatial Relationships

1 Introduction

The development of text-to-image (T2I) diffusion models such as Stable Diffusion [49] and DALL-E 3 [39] has led to the growth of image synthesis frameworks that

⋆ Equal contribution. Correspondence to agneet@asu.edu

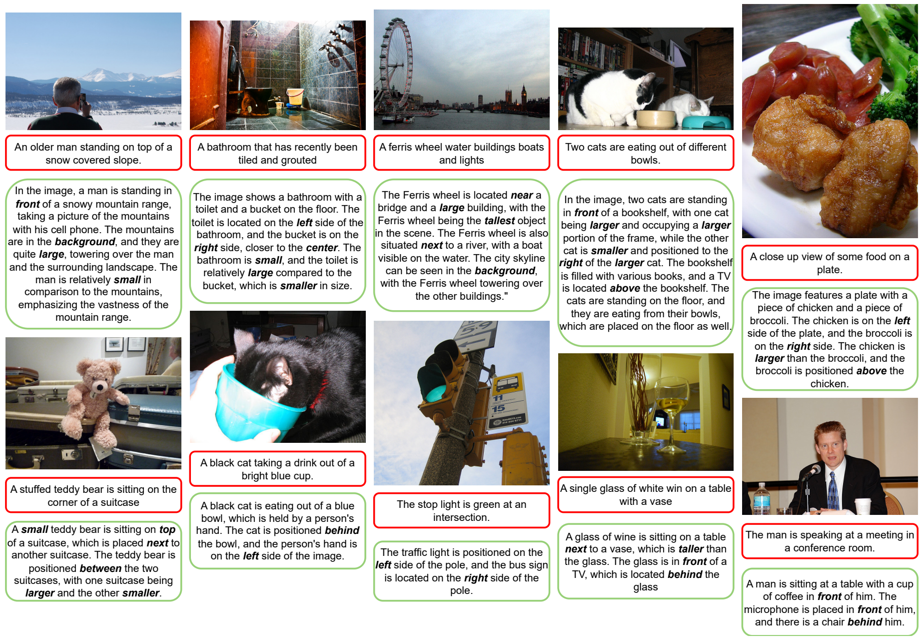

Fig. 1: We find that existing vision-language datasets do not capture spatial relationships well. To alleviate this shortcoming, we synthetically re-caption ∼6M images with a spatial focus, and create the SPRIGHT (**SP**atially **RIGHT**) dataset. Shown above are samples from the COCO Validation Set, where text in red denotes ground-truth captions and text in green are corresponding captions from SPRIGHT.

are able to generate high resolution photo-realistic images. These models have been adopted widely in downstream applications such as video generation [54], image editing [20], robotics [15], and more. Multiple variations of T2I models have also been developed, which vary according to their text encoder [5], priors [47], and inference efficiency [36]. However, a common bottleneck that affects all of these methods is their inability to generate spatially consistent images: that is, given a natural language prompt that describes a spatial relationship, these models are unable to generate images that faithfully adhere to it.

In this paper, we present a holistic approach towards investigating and mitigating this shortcoming through diverse lenses. We develop datasets, efficient training techniques, and explore multiple ablations and analyses to understand the behaviour of T2I models towards prompts that contain spatial relationships.

Our first finding reveals that existing vision-language (VL) datasets lack sufficient representation of spatial relationships. Although frequently used in the English lexicon, we find that spatial words are scarcely found within image-text pairs of the existing datasets. To alleviate this shortcoming, we create the "SPRIGHT" (**SP**atially **RIGHT**) dataset, the first spatially-focused large scale dataset. Specifically, we synthetically re-caption ∼6 million images sourced from

4 widely used datasets, with a spatial focus (Section 3). As shown in Figure 1, SPRIGHT captions describe the fine-grained relational and spatial characteristics of an image, whereas human-written ground truth captions fail to do so. Through a 3-fold comprehensive evaluation and analysis of the generated captions, we benchmark the quality of the generated captions and find that SPRIGHT largely improves over existing datasets in its ability to capture spatial relationships. Next, leveraging only ∼0.25% of our dataset, we achieve a 22% improvement on the T2I-CompBench [22] Spatial Score, and a 31.04% and 29.72% improvement in the FID [21] and CMMD scores [23], respectively.

Our second finding reveals that significant performance improvements in spatial consistency of a T2I model can be achieved by fine-tuning on images that contain a large number of objects. We achieve *state-of-the-art* performance, and improve image fidelity, by fine-tuning on *<500* image-caption pairs from SPRIGHT; training only on images that have a large number of objects. As investigated in VISOR [19], models often fail to generate the mentioned objects in a spatial prompt; we posit that by optimizing the model over images which have a large number of objects (and consequently, spatial relationships), we teach it to generate a large number of objects, which positively impacts its spatial consistency. In addition to improving spatial consistency, our model achieves large gains in performance across all aspects of T2I generation; generating correct number of distinct objects, attribute binding and accurate generation in response to complex prompts.

We further demonstrate the impact of SPRIGHT by benchmarking the trade-offs achieved with long and short spatial captions, as well as spatially focused and general captions. We take the first steps towards discovering layer-wise activation patterns associated with spatial relationships, by examining the representation space of CLIP [45] as a text encoder.

Our contributions and key findings are summarized below:

– We create SPRIGHT, the first spatially focused, large scale vision-language dataset by re-captioning ∼6 million images from 4 widely used existing datasets.To demonstrate the efficacy of SPRIGHT, we fine-tune baseline Stable Diffusion models on a small subset of our data and achieve performance gains across multiple spatial reasoning benchmarks while improving the corresponding FID and CMMD scores.
– We achieve state-of-the-art performance on spatial relationships by developing an efficient training methodology; specifically, we optimize over a small number (<500) of images which consists of a large number of objects, and achieve a 41% improvement over our baseline model.
– Through multiple ablations and analyses, we present our findings related to spatial relationships: the impact of long captions, the trade-off between spatial and general captions, layer-wise activations of the CLIP text encoder, effect of training with negations and improvements over attention maps.

2 Related Work

Text-to-image generative models. Since the initial release of Stable Diffusion [49] and DALL-E [48], different classes of T2I models have been developed, all optimized to generate highly realistic images corresponding to complex natural language prompts. Models such as PixArt-Alpha [5], Imagen [50], and ParaDiffusion [55] move away from the CLIP text encoder, and explore traditional language models such as T5 [46] and LLaMA [53] to process text prompts. unCLIP [47] based models have led to multiple methods [29,42] that leverage a CLIP-based prior as part of their diffusion pipeline.

Spatial relationships in T2I models. Benchmarking the failures of T2I models on spatial relationships has been well explored by VISOR [19], T2I-CompBench [22], GenEval [16], and DALL-E Eval [8]. Both training-based and test-time adaptations have been developed to specifically improve upon these benchmarks. Control-GPT [61] finetunes a ControlNet [60] model by generating TikZ code representations with GPT-4 and optimizing over grounding tokens to generate images. SpaText [1], GLIGEN [30], and ReCo [57] are training-based methods that introduce additional conditioning in their fine-tuning process to achieve better spatial control for image generation. LLM-Grounded Diffusion [31] is a test-time multi-step method that improves over layout generated LLMs in an iterative manner. Layout Guidance [6] restricts objects to their annotated bounding box locations through refinement of attention maps during inference. LayoutGPT [14] creates an LLM guided initial layout in the form of CSS, and then uses layout-to-image models to create indoor scenes.

Synthetic captions for T2I models. The efficacy of using descriptive and detailed captions has recently been explored by DALL-E 3 [39], PixArt-Alpha [5] and RECAP [52]. DALL-E 3 builds an image captioning module by jointly optimizing over a CLIP and language modeling objective. RECAP fine-tunes an image captioning model (PALI [7]) and reports the advantages of fine-tuning the Stable Diffusion family of models on long, synthetic captions. PixArt-Alpha also re-captions images from the LAION [51] and Segment Anything [25] datasets; however their key focus is to develop descriptive image captions. On the contrary, our goal is to develop captions that explicitly capture the spatial relationships seen in the image.

3 The SPRIGHT Dataset

We find that current vision-language (VL) datasets do not contain "enough" relational and spatial relationships. Despite being frequently used in the English vocabulary [6], words like "left/right", "above/behind" are scarce in existing VL datasets. This holds for both annotator-provided captions, e.g., COCO [32], and web-scraped alt-text captions, e.g., LAION [51]. We posit that the absence of

[6] `https://www.oxfordlearnersdictionaries.com/us/wordlists/oxford3000-5000`

such phrases is one of the fundamental reasons for the lack of spatial consistency in current text-to-image models. Furthermore, language guidance is now being used to perform mid-level [56,59] and low-level [26,63] computer vision tasks. This motivates us to create the SPRIGHT (**SP**atially **RIGHT**) dataset, which explicitly encodes fine-grained relational and spatial information found in images.

3.1 Creating the SPRIGHT Dataset

We re-caption approximately six million images from four existing vision-language datasets, *i.e.* datasets containing images and their corresponding natural language descriptions:

- **CC-12M** [3] : We re-caption a total of 2.3 million images from the CC-12M dataset, filtering out images of resolution less than 768 × 768.
- **Segment Anything (SA)** [25] : We select Segment Anything as most images in it encapsulates a large number of objects; i.e. larger number of spatial relationships can be captured from a given image. We re-caption 3.5 million images as part of our re-captioning process. Since SA does not have ground-truth captions, we generate its *general* captions using the CoCa [58] model.
- **COCO** [32] : We re-caption images (∼ 40,000) from the validation set.
- **LAION-Aesthetics**[7] : We used 50,000 images from LAION-Aesthetics. [8]

We use LLaVA-1.5-13B [33] with the following prompt to produce synthetic *spatial captions* to create the SPRIGHT dataset:

```
Using 2 sentences, describe the spatial relationships seen in the image. You can use
words like left/right, above/below, front/behind, far/near/adjacent, inside/outside.
Also describe relative sizes of objects seen in the image.
```

3.2 Impact of SPRIGHT

Table 1 shows that SPRIGHT enhances the presence of spatial phrases across all relationship types on all the datasets. For 11 relationships, while the ground-truth captions of COCO and LAION only capture 21.05% and 6.03% of relationships, SPRIGHT captures 304.79% and 284.7%, respectively, *i.e.* each re-captioned COCO image in SPRIGHT has ∼3 spatial phrases. This shows that captions in VL datasets largely lack the presence of spatial relationships, and that SPRIGHT is able to improve upon this shortcoming by almost always capturing spatial relationships in every sentence. Our captions offer several improvements beyond the spatial aspects: **(i)** As depicted in Table 2 we improve the overall linguistic quality compared to the original captions, and **(ii)** we identify more objects and amplify their occurrences as illustrated in Figure 2; where we plot the top 10 objects present in the original COCO Captions and find that we significantly upsample their corresponding presence in SPRIGHT.

[7] `https://laion.ai/blog/laion-aesthetics/`

[8] The entire LAION-5B dataset has been recalled for safety review: `https://laion.ai/notes/laion-maintenance/`. We will release our re-captioning outputs for these images based on the conclusions of this safety review.

Table 1: Compared to ground truth annotations, SPRIGHT consistently improves the presence of relational and spatial relationships captured in its captions, across diverse images from different datasets.

Dataset	**% of Spatial Phrases**										
	left	right	above	below	front	behind	next	close	far	small	large
COCO	0.16	0.47	0.61	0.15	3.39	1.09	6.17	1.39	0.19	3.28	4.15
+ SPRIGHT	26.80	23.48	21.25	5.93	41.68	21.13	36.98	15.85	1.34	48.55	61.80
CC-12M	0.61	1.45	0.40	0.19	1.40	0.43	0.54	0.94	1.07	1.44	1.44
+ SPRIGHT	24.53	22.36	20.42	6.48	41.23	14.37	22.59	12.9	1.10	43.49	66.74
LAION	0.27	0.75	0.16	0.05	0.83	0.11	0.24	0.67	0.91	1.03	1.01
+ SPRIGHT	24.36	21.7	14.27	4.07	42.92	16.38	26.93	13.05	1.16	49.59	70.27
Segment Anything	0.02	0.07	0.27	0.06	5.79	0.19	3.24	7.51	0.05	0.85	10.58
+ SPRIGHT	18.48	15.09	23.75	6.56	43.5	13.58	33.02	11.9	1.25	52.19	80.22

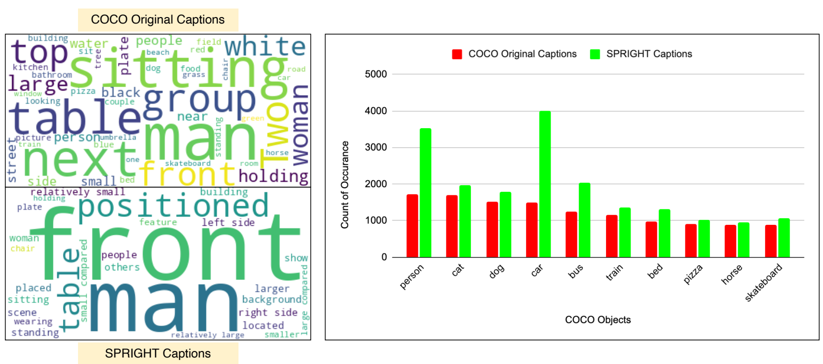

Fig. 2: Compared to ground truth COCO captions,(**Left**) Word cloud representations showing that SPRIGHT captions significantly amplify the presence of spatial relationships. (**Right**)SPRIGHT captions also capture a higher number of object occurances.

3.3 Dataset Validation

We perform 3 levels of evaluation to validate the SPRIGHT captions:

1. FAITHScore. Following [24], we leverage a large language model to deconstruct generated captions into atomic (simple) claims that can be individually and independently verified in a Visual Question Answering (VQA) format. We randomly sample 40,000 image-generated caption pairs from our dataset, and prompt GPT-3.5-Turbo to identify descriptive phrases (as opposed to subjective analysis that cannot be verified from the image) and decompose the descriptions into atomic statements. These atomic statements are then passed to LLaVA-1.5-13B for verification, and correctness is aggregated over 5 categories: entity, relation, colors, counting, and other attributes. We also measure correctness on spatial-related atomic statements, i.e., those containing one of the keywords

Table 2: In addition to improving the presence of spatial relationships, SPRIGHT enhances linguistic diversity of captions in comparison to their original versions.

Dataset	**Average / caption**			
	Nouns	Adjectives	Verbs	Tokens
COCO → COCO+SPRIGHT	3.00 → 14.31	0.83 → 3.82	0.04 → 0.15	11.28 → 68.22
CC-12M → CC-12M+SPRIGHT	3.35 → 13.99	1.36 → 4.36	0.26 → 0.16	22.93 → 67.41
LAION → LAION+SPRIGHT	1.78 → 14.32	0.70 → 4.53	0.11 → 0.14	12.49 → 69.74
SA → SA+SPRIGHT	3.10 → 13.42	0.79 → 4.65	0.01 → 0.12	09.88 → 63.90

left/right, above/below, near/far, large/small and background/foreground. The captions are on average 88.9% correct, with spatially-focused relations, being 83.6% correct; with the detailed breakdown presented in the Supplementary Materials. Since there is some uncertainty about bias induced by using LLaVA to evaluate LLaVA-generated captions, we also verify the caption quality in other ways, as described next.

2. GPT-4 (V). Inspired by recent methods [39, 64], we perform a small-scale study on a split of 444 images from LAION and SA (from Section 4.2) to evaluate our captions with GPT-4(V) Turbo [40]. We prompt GPT-4(V) to rate each caption between a score of 1 to 10, especially focusing on the correctness of the spatial relationships captured. Captions of images from LAION and SA had a `{mean, median}` rating of `{7.49,8}` and `{7.36,8}`, respectively. We present the prompt used in the Supplementary Materials.

3. Human Annotation. We also annotate a total of 3,000 images through a crowd-sourced human study, where each participant annotates a maximum of 30 image-text pairs. As evidenced by the average number of tokens in Table 1, most captions in SPRIGHT have >1 sentences. Therefore, for fine-grained evaluation, we randomly select 1 sentence, from a caption in SPRIGHT, and evaluate its correctness for a given image. Across 149 responses, we find the metrics to be: `correct=1840` and `incorrect=928`, yielding an accuracy of 66.57%.

4 Improving Spatial Consistency

In this section, we leverage SPRIGHT in an effective and efficient manner, and describe methodologies that significantly advance spatial reasoning in T2I models. We use Stable Diffusion v2.1 [9] as the base model and our training and validation set consists of 13,500 and 1,500 images respectively, randomly sampled in a 50:50 split between LAION-Aesthetics and Segment Anything. Each image is paired with a typical caption and a spatial caption (from SPRIGHT). During fine-tuning, for each image, we randomly choose one of the given caption types in a 50:50 ratio. We fine-tune the U-Net and the CLIP text encoder as part

[9] `https://huggingface.co/stabilityai/stable-diffusion-2-1`

Table 3: Quantitative metrics across multiple spatial reasoning and image fidelity metrics, demonstrating the effectiveness of high quality spatially-focused captions in SPRIGHT. Green indicates results of the model fine-tuned on SPRIGHT. For FID, we use cfg = 3.0 and 7.0 for the baseline and the fine-tuned model, respectively.

Method	OA (%) (↑)	VISOR (%) (↑)						T2I-CompBench (↑) Spatial Score	ZS-FID (↓)	CMMD (↓)
		uncond	cond	1	2	3	4			
SD 2.1	47.83	30.25	63.24	64.42	35.74	16.13	4.70	0.1507	21.646	0.703
+ SPRIGHT	53.59	36.00	67.16	66.09	44.02	24.15	9.13	0.1840	14.925	0.494

Table 4: Across all reported methods, we achieve *state-of-the-art* performance on the T2I-CompBench Spatial Score. This is achieved by fine-tuning SD 2.1 on 444 image-caption pairs from the SPRIGHT dataset; where each image has >18 objects.

# of Objects per Image	<6	<11	11	>11	> 18
# of Training Images	444	1346	1346	1346	444
T2I-CompBench Spatial Score (↑)	0.1309	0.1468	0.1667	0.1613	**0.2133**

of our training, both with a learning rate 5×10^{-6} optimized by AdamW [35] and a global batch size of 128. While we train the U-Net for 15,000 steps, the CLIP text encoder remains frozen during the first 10,000 steps. We develop our code-base on top of the Diffusers library [43].

4.1 Improving upon Baseline Methods

We present results on the spatial relationship benchmarks (VISOR [19], T2I-CompBench [22]) and image fidelity metrics in Table 3. To account for the inconsistencies associated with FID [9,41], we also report results on CMMD [23]. Across all metrics, our method significantly improves upon the base model by fine-tuning on <***15k*** images. We conclude that the dense, spatially focused captions in SPRIGHT provide effective spatial guidance to T2I models, and alleviate the need to scale up fine-tuning on a large number of images. As shown in Figure 3, the model captures complex spatial relationships (`top right`), relative sizes (`large`) and patterns (`swirling`).

4.2 Efficient Training Methodology

We devise an additional efficient training methodology, which achieves state-of-the-art performance on the spatial aspect of the T2I-CompBench Benchmark. We hypothesize that ***(a)*** images that capture a large number of objects inherently also contain multiple spatial relationships; and ***(b)*** training on these kinds of images will optimize the model to consistently generate a large number of objects, given a prompt containing spatial relationships; a well-documented failure mode of current T2I models [19].

For our dataset of <**15k** images the median # of objects/image = *11*. We partition our dataset into multiple subsets based on the maximum number of

Fig. 3: Generated images from our model, as described in Section 4.1, on prompts which contain multiple objects and complex spatial relationships. We curate these prompts from ChatGPT.

objects present in an image. This partitioning is automated using the open-world image tagging model Recognize Anything [62]. We create five subsets, train corresponding models on a single subset and benchmark them in Table 4. We keep the same hyper-parameters as before, only initiating training of the CLIP Text Encoder from the beginning. With an increase in the # of objects / image, iterative improvement in spatial fidelity is observed, with the best score for the subset containing greater than 18 objects.

Our major finding is that, with **444** training images and spatial captions from SPRIGHT, we achieve a 41% improvement over the baseline SD 2.1 and attain state-of-the-art performance across all reported models on the T2I-CompBench spatial score. In Table 5, compared to SD 2.1, we significantly improve all aspects of the VISOR score, while also enhancing the ZS-FID and CMMD scores on COCO-30K images by 25.39% and 27.16%, respectively. Our key findings on VISOR (Table 6) include: (a) a 26.86% increase in the Object Accuracy (OA) score, indicating substantial gains in generating objects mentioned in the input prompt, and (b) a VISOR_4 score of 16.15%, demonstrating our model's consistent generation of spatially accurate images.

We also compare our model's performance on the GenEval [16] benchmark (Table 7), and find that in addition to improving spatial relationship (see *Posi-*

Table 5: Comparing baseline SD 2.1 with our state-of-the-art model, across multiple spatial reasoning and image fidelity metrics, as described in Section 4.2. Green indicates results from our model. For FID, we use cfg = 3.0 and 7.5 for the baseline model and our model, respectively

Method	OA (%) (↑)	VISOR (%) (↑)						T2I-CompBench (↑) Spatial Score	ZS-FID (↓)	CMMD (↓)
		uncond	cond	1	2	3	4			
SD 2.1	47.83	30.25	63.24	64.42	35.74	16.13	4.70	0.1507	21.646	0.703
+ SPRIGHT (<500 images)	60.68	43.23	71.24	71.78	51.88	33.09	16.15	0.2133	16.149	0.512

Table 6: Results on the VISOR Benchmark. Our model outperforms existing methods, on all aspects related to spatial relationships, consistently generating spatially accurate images as shown by the high VISOR [1-4] values.

Method	OA (%)	VISOR (%)					
		uncond	cond	1	2	3	4
GLIDE [38]	3.36	1.98	59.06	6.72	1.02	0.17	0.03
GLIDE + CDM [34]	10.17	6.43	63.21	20.07	4.69	0.83	0.11
CogView2 [11]	18.47	12.17	<u>65.89</u>	33.47	11.43	3.22	0.57
DALLE-mini [10]	27.10	16.17	59.67	38.31	17.50	6.89	1.96
DALLE-2 [47]	**63.93**	<u>37.89</u>	59.27	**73.59**	<u>47.23</u>	<u>23.26</u>	<u>7.49</u>
Structured Diffusion [13]	28.65	17.87	62.36	44.70	18.73	6.57	1.46
Attend-and-Excite [4]	42.07	25.75	61.21	49.29	19.33	4.56	0.08
Ours (<500 images)	<u>60.68</u>	**43.23**	**71.24**	<u>71.78</u>	**51.88**	**33.09**	**16.15**

tion), our model shows improvement in generating 1 and 2 objects, along with the correct number of objects. Throughout our experiments, our training approach not only preserves but also enhances the *non-spatial* aspects associated with a text-to-image model. Additional results and illustrations from VISOR and T2I-CompBench are provided in the Supplementary Materials.

5 Ablation Studies and Analyses

To fully ascertain the impact of spatially-focused captions in SPRIGHT, we experiment with multiple nuances of our dataset and the corresponding T2I pipeline. Unless stated otherwise, the experimental setup identical to Section 4.

5.1 Optimal Ratio of Spatial Captions

To understand the impact of spatially focused captions in comparison to ground-truth captions, we fine-tune different models by varying the % of spatial captions. The results suggest that the model trained on 50% spatial captions achieves the best spatial scores on T2I-CompBench (Table 8 (a)). The models trained on only 25% of spatial captions suffer largely from incorrect spatial relationships whereas

Table 7: Results on the GenEval Benchmark. In addition to spatial relationships, we also improve model performance in generating the correct number of objects.

Method	**Overall**	**Single object**	**Two objects**	**Counting**	**Colors**	**Position**	**Attribute binding**
CLIP retrieval [2]	0.35	0.89	0.22	0.37	0.62	0.03	0.00
minDALL-E [28]	0.23	0.73	0.11	0.12	0.37	0.02	0.01
SD 1.5	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD 2.1	0.50	0.98	0.51	0.44	**0.85**	0.07	0.17
SDXL [44]	**0.55**	0.98	**0.74**	0.39	**0.85**	**0.15**	**0.23**
PixArt-Alpha [5]	0.48	0.98	0.50	0.44	0.80	0.08	0.07
Ours (<500 images)	0.51	**0.99**	0.59	**0.49**	**0.85**	0.11	0.15

Table 8: Comparing (a) the effect the percentage of spatial captions and (b) the effect of long and short spatial captions.

% of spatial captions	**T2I-CompBench Spatial Score** (↑)
25	0.154
50	0.178
75	0.161
100	0.140

(a) T2I-CompBench Spatial Scores for models trained on varying ratios of spatial captions. Fine-tuning on a ratio of 50% and 75% of spatial captions yields optimal results.

Model, Setup	**T2I-CompBench Spatial Score (↑)** Long Captions	Short Captions
SD 1.5, w/o CLIP FT	0.0910	0.0708
SD 2.1, w/o CLIP FT	0.1605	0.1420
SD 2.1, w/ CLIP FT	0.1777	0.1230

(b) T2I-CompBench Spatial Scores for models trained on long and short *spatial* captions. Across multiple setups, we find that longer spatial captions lead to better improvements in spatial consistency.

the model trained only on spatial captions fails to generate the mentioned objects in the input prompt. Figure 4 shows illustrative examples.

5.2 Impact of Long and Short Spatial Captions

We also compare the effect of fine-tuning with shorter and longer variants of spatial captions. We create the shorter variants by randomly sampling 1 sentence from the longer caption, and fine-tune multiple models, with different setups. Across, all setups, (Table 8 (b)) longer captions perform better than their shorter counterparts. In fact, CLIP fine-tuning hurts performance while using shorter captions, but has a positive impact on longer captions. This potentially happens because fine-tuning CLIP enables T2I models to generalize better to longer captions, which are out-of-distribution at the onset of training as they are initially pre-trained on short(er) captions from datasets such as LAION.

5.3 Investigating the CLIP Text Encoder

The CLIP Text Encoder enables semantic understanding of the input text prompts in the Stable Diffusion model. As we fine-tune CLIP on the spatial captions, we investigate the various nuances associated with it:

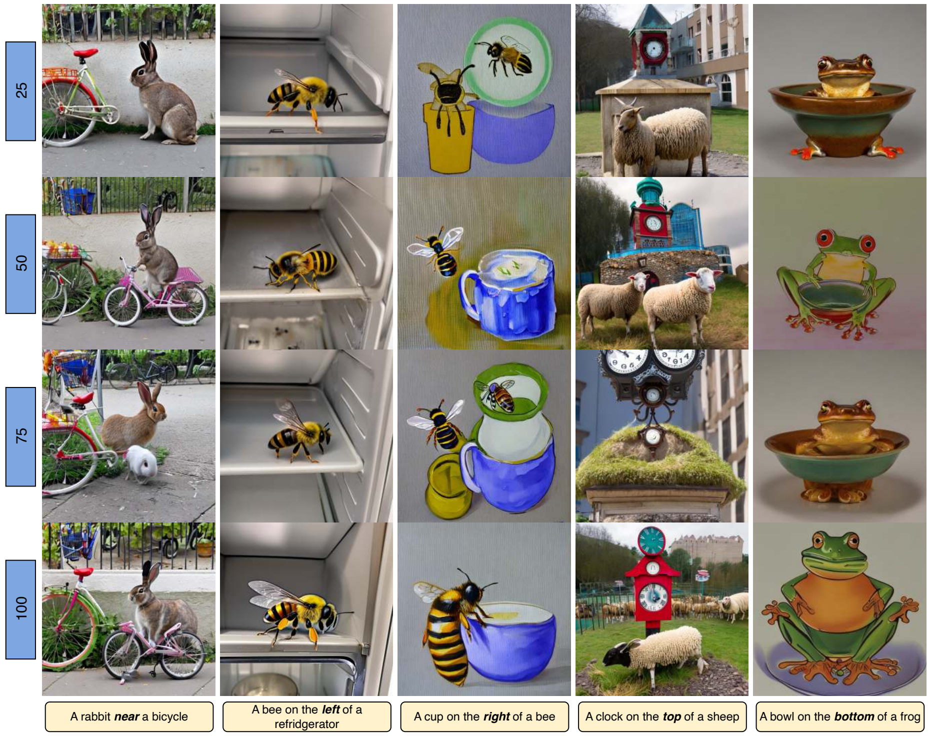

Fig. 4: Illustrative comparisons between models trained on varying ratio of spatial experiments. Models trained on 50% and 75% spatial captions are optimal.

Centered Kernel Alignment (CKA) [27,37] compares layer-wise representations learned by two neural networks. Figure 5 illustrates different representations learned by baseline CLIP, compared against the one trained on SPRIGHT. We compare layer activations across 50 simple and complex prompts and aggregate representations from all the layers. Our findings reveal that the MLP and output attention projection layers play a larger role in enhancing spatial comprehension, as opposed to layers such as the layer norm. This distinction is larger with complex prompts, showing that the longer prompts from SPRIGHT indeed lead to more diverse embeddings being learned within the CLIP space.

Improving Semantic Understanding : To evaluate semantic understanding of the fine-tuned CLIP, we perform the following experiment: given a prompt containing a spatial phrase and 2 objects, we modify the prompt by switching the objects (*e.g.* "an airplane above an apple" → "an apple above an airplane"). Although these sentences have the same words, the placement of the two nouns relative to the preposition "above" completely changes the meaning of the sentence. To evaluate if models can discern this spatial distinction, we compute the cosine similarity between the pooled layer outputs of the original and modified

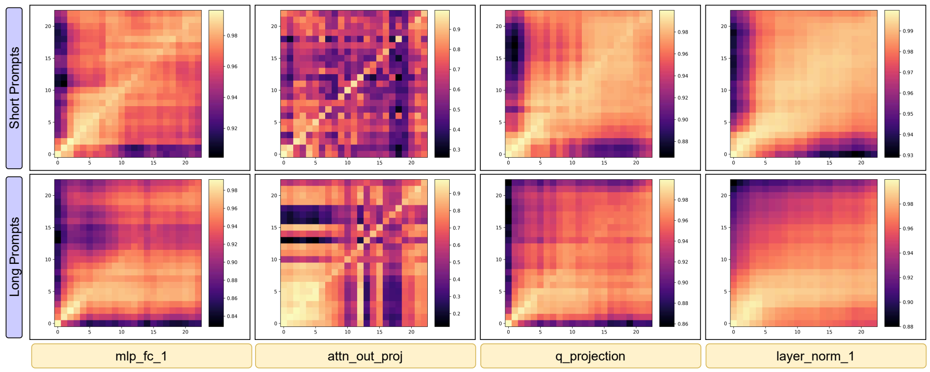

Fig. 5: Comparison of layer-wise representations between Baseline CLIP (X-axis) and fine-tuned CLIP on SPRIGHT (Y-axis). Spatial captions show distinct representations in output attention projections and MLP layers, while layer norm layers are more similar. The representation gap widens with long, complex prompts, suggesting spatial prompts in SPRIGHT create diverse embeddings.

Table 9: CLIP fine-tuned on SPRIGHT is able to differentiate the spatial nuances present in a textual prompt. While Baseline CLIP shows a high similarity for ***spatially different*** prompts, SPRIGHT enables better fine-grained understanding.

	"above"	"below"	"to the left of"	"to the right of"	"in front of"	"behind"
Baseline CLIP	0.9225	0.9259	0.9229	0.9223	0.9231	0.9289
CLIP + SPRIGHT	0.8674	0.8673	0.8658	0.8528	0.8417	0.8713

prompts, for $\sim 37k$ sentences. Table 9 shows that CLIP finetuned on SPRIGHT is able to differentiate between the prompts better (*i.e.* lower cosine similarity) than the baseline.

5.4 Improvement over Attention Maps

Inspired by methods like Attend-and-Excite [4], we visualize attention relevancy maps for both simple and complex spatial prompts. Our model better generates the expected objects and achieves improved spatial localization compared to the baseline. For instance, the baseline models fails to generate objects like the bed and house, which our model successfully generates. The relevancy map indicates that high attention patches for missing words are spread across the image. Additionally, our model correctly attends to spatial words in the image, unlike the baseline. For example, in our model (Figure 6, bottom row), `below` attends to patches below the bed, and `right` attends to patches on the road's right, while Stable Diffusion 2.1 does not. We achieve these improvements across the intermediate attention maps and the final generated images.

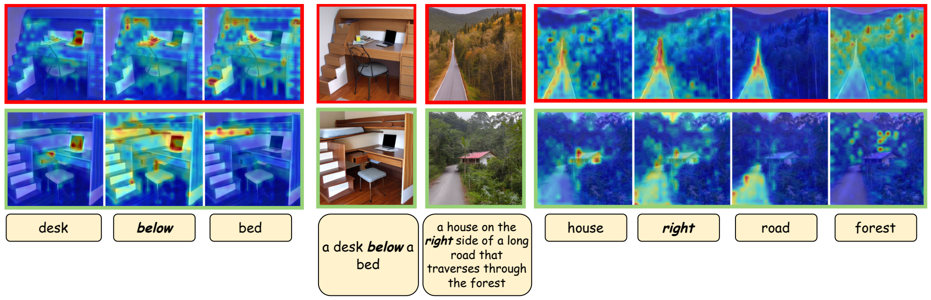

Fig. 6: Visualising the cross-attention relevancy maps for baseline (**top row**) and fine-tuned model (**bottom row**) on SPRIGHT. Images in red are from baseline model while images in green are from our model.

5.5 Training with Negation

Dealing with negation remains a challenge for multimodal models as reported by previous findings on Visual Question Answering and Reasoning [12, 17, 18]. Thus, in this section, we investigate the ability of T2I models to reason over spatial relationships and negations, simultaneously. Specifically, we study the impact of training a model with `‘‘A man is not to the left of a dog’’` as a substitute to `‘‘A man is to the right of a dog’’`. To create such captions, we post-process our generated captions and randomly replace spatial occurrences with their *negation* counter-parts, and ensure that the semantic meaning of the sentence remains unchanged. Training on such a model, we find slight improvements in the spatial score, both while evaluating on prompts containing only negation (***0.069*** $>$ ***0.066***) and those that contain a mix of negation and simple statements (***0.1427*** $>$ ***0.1376***). There is however, a significant drop in performance, when evaluating on prompts that only contain negation; thus highlighting a major scope of improvement in this regard.

6 Conclusion

In this work, we present findings and techniques that enable improvement of spatial relationships in text-to-image models. We develop a large-scale dataset, SPRIGHT that captures fine-grained spatial relationships across a diverse set of images. Leveraging SPRIGHT, we develop efficient training techniques and achieve state-of-the art performance in generating spatially accurate images. We thoroughly explore various aspects concerning spatial relationships and evaluate the range of diversity introduced by the SPRIGHT dataset. We leave further scaling studies related to spatial consistency as future work. We believe our findings and results facilitate a comprehensive understanding of the interplay between spatial relationships and T2I models, and contribute to the future development of robust vision-language models.

Acknowledgements

We thank Lucain Pouget for helping us in uploading the dataset to the Hugging Face Hub and the Hugging Face team for providing computing resources to host our demo. The authors acknowledge resources and support from the Research Computing facilities at Arizona State University. AC, CB, YY were supported by NSF Robust Intelligence program grants #1750082 and #2132724. TG was supported by Microsoft's Accelerating Foundation Model Research (AFMR) program and UMBC's Strategic Award for Research Transitions (START). The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

1. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.01762, http://dx.doi.org/10.1109/CVPR52729.2023.01762
2. Beaumont, R.: Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval (2022)
3. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
4. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
5. Chen, J., Jincheng, Y., Chongjian, G., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: The Twelfth International Conference on Learning Representations (2023)
6. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024)
7. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
8. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3043–3054 (2023)
9. Chong, M.J., Forsyth, D.: Effectively unbiased fid and inception score and where to find them. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6070–6079 (2020)
10. Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Le Khac, P., Melas, L., Ghosh, R.: Dall·e mini (7 2021). https://doi.org/10.5281/zenodo.5146400, https://github.com/borisdayma/dalle-mini

11. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. Advances in Neural Information Processing Systems **35**, 16890–16902 (2022)
12. Dobreva, R., Keller, F.: Investigating negation in pre-trained vision-and-language models. In: Bastings, J., Belinkov, Y., Dupoux, E., Giulianelli, M., Hupkes, D., Pinter, Y., Sajjad, H. (eds.) Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. pp. 350–362. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.blackboxnlp-1.27, https://aclanthology.org/2021.blackboxnlp-1.27
13. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis (2023)
14. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. Advances in Neural Information Processing Systems **36** (2024)
15. Gao, J., Hu, K., Xu, G., Xu, H.: Can pre-trained text-to-image models generate visual goals for reinforcement learning? Advances in Neural Information Processing Systems **36** (2024)
16. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=Wbr51vK331
17. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Vqa-lol: Visual question answering under the lens of logic. In: European conference on computer vision. pp. 379–396. Springer (2020)
18. Gokhale, T., Chaudhary, A., Banerjee, P., Baral, C., Yang, Y.: Semantically distributed robust optimization for vision-and-language inference. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 1493–1513. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.118, https://aclanthology.org/2022.findings-acl.118
19. Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
22. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
23. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9307–9315 (2024)
24. Jing, L., Li, R., Chen, Y., Jia, M., Du, X.: Faithscore: Evaluating hallucinations in large vision-language models. arXiv preprint arXiv:2311.01477 (2023)

25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
26. Kondapaneni, N., Marks, M., Knott, M., Guimaraes, R., Perona, P.: Text-image alignment for diffusion-based perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13883–13893 (2024)
27. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMLR (2019)
28. kuprel: min-dalle (2022), https://github.com/kuprel/min-dalle
29. Lee, D., Kim, J., Choi, J., Kim, J., Byeon, M., Baek, W., Kim, S.: Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo (2022)
30. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
31. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. ArXiv preprint **abs/2305.13655** (2023), https://arxiv.org/abs/2305.13655
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
33. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
34. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
36. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023)
37. Nguyen, T., Raghu, M., Kornblith, S.: Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth (2021)
38. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models (2022)
39. OpenAI: Dalle-3 (2023), https://openai.com/dall-e-3
40. OpenAI: Gpt-4(v) (2023), https://cdn.openai.com/papers/GPTV_System_Card.pdf
41. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022)
42. Patel, M., Kim, C., Cheng, S., Baral, C., Yang, Y.: Eclipse: A resource-efficient text-to-image prior for image generations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9069–9078 (2024)

43. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)
44. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
46. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
47. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
48. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (2021), http://proceedings.mlr.press/v139/ramesh21a.html
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
50. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
51. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
52. Segalis, E., Valevski, D., Lumen, D., Matias, Y., Leviathan, Y.: A picture is worth a thousand words: Principled recaptioning improves image generation. arXiv preprint arXiv:2310.16656 (2023)
53. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang,

Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
54. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
55. Wu, W., Li, Z., He, Y., Shou, M.Z., Shen, C., Cheng, L., Li, Y., Gao, T., Zhang, D., Wang, Z.: Paragraph-to-image generation with information-enriched diffusion model. arXiv preprint arXiv:2311.14284 (2023)
56. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2945–2954 (June 2023)
57. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255 (2023)
58. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022), https://openreview.net/forum?id=Ee277P3AYC
59. Yun, S., Park, S.H., Seo, P.H., Shin, J.: Ifseg: Image-free semantic segmentation via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2967–2977 (June 2023)
60. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
61. Zhang, T., Zhang, Y., Vineet, V., Joshi, N., Wang, X.: Controllable text-to-image generation with gpt-4. arXiv preprint arXiv:2305.18583 (2023)
62. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1724–1732 (2024)
63. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5729–5739 (2023)
64. Zhong, M., Shen, Y., Wang, S., Lu, Y., Jiao, Y., Ouyang, S., Yu, D., Han, J., Chen, W.: Multi-lora composition for image generation. arXiv preprint arXiv:2402.16843 (2024)