DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models

Jaemin Cho Abhay Zala Mohit Bansal
UNC Chapel Hill
{jmincho, aszala, mbansal}@cs.unc.edu

Abstract

Recently, DALL-E [45], a multimodal transformer language model, and its variants including diffusion models have shown high-quality text-to-image generation capabilities. However, despite the realistic image generation results, there has not been a detailed analysis of how to evaluate such models. In this work, we investigate the visual reasoning capabilities and social biases of different text-to-image models, covering both multimodal transformer language models and diffusion models. First, we measure three visual reasoning skills: object recognition, object counting, and spatial relation understanding. For this, we propose PAINTSKILLS*, a compositional diagnostic evaluation dataset that measures these skills. Despite the high-fidelity image generation capability, a large gap exists between the performance of recent models and the upper bound accuracy in object counting and spatial relation understanding skills. Second, we assess the gender and skin tone biases by measuring the gender/skin tone distribution of generated images across various professions and attributes. We demonstrate that recent text-to-image generation models learn specific biases about gender and skin tone from web image-text pairs. We hope our work will help guide future progress in improving text-to-image generation models on visual reasoning skills and learning socially unbiased representations.*[1]

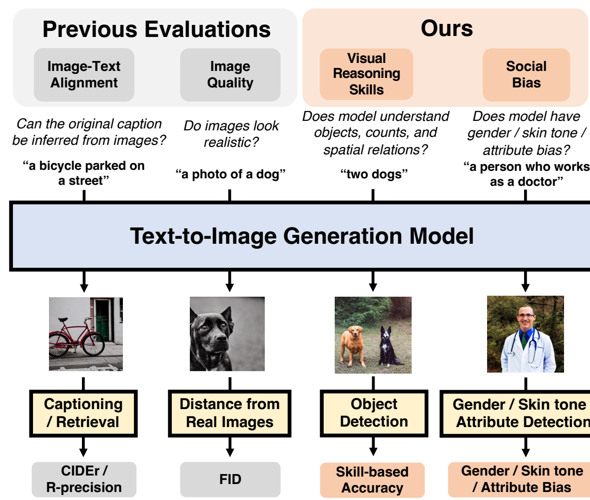


Figure 1. Overview of our proposed evaluation process for text-to-image generation models. In addition to conventional image-text alignment and image quality evaluation, we propose to measure visual reasoning skills (Sec. 4.1) and social biases (Sec. 4.2) of models. The example images are generated with Stable Diffusion.

1. Introduction

Generating images from textual descriptions based on machine learning is an active research area [21]. Recently, DALL-E [45], a 12B parameter transformer [60] trained to generate images from text, has shown a diverse set of generation capabilities, including creating anthropomorphic objects, editing images, and rendering text, which previous models have never shown. Even though DALL-E and its variants have gained much attention, there has not been a concrete quantitative analysis of what they can do.

Most works have only evaluated their text-to-image generation models with two types of automated metrics [21]: 1) image-text alignment [69, 30, 26] - whether the generated images align with the semantics of the text descriptions; 2) image quality [52, 25] - whether the generated images look similar to images from training data. Hence, to provide novel insights into the abilities and limitations of text-to-image generation models, we propose to evaluate their **visual reasoning skills** and **social biases**, in addition to the previously proposed image-text alignment and image quality metrics. Since the original DALL-E checkpoint is not available, in our experiments, we choose four popular text-to-image generation models that publicly release their code and checkpoints: DALL-E$^{\text{Small}}$ [64], minDALL-E [33], Stable Diffusion [49], and Karlo [35].

First, we introduce PAINTSKILLS, a compositional di-

[1] Code and data: https://github.com/j-min/DallEval

agnostic evaluation dataset that measures three fundamental visual reasoning capabilities: object recognition, object counting, and spatial relation understanding. To avoid statistical bias that hinders models from learning compositional reasoning [23, 1, 15, 17], for PAINTSKILLS, we create images based on a 3D simulator and control our images to have a uniform distribution over objects and relations. To calculate the score for each skill, we employ a widely-used DETR object detector [11] on the PAINTSKILLS dataset that can detect objects on the test split images with very high oracle accuracy. We also show that our object detection-based evaluation is highly correlated with human judgment. Then we measure whether the objects in the images satisfy the skill-specific semantics of the input text (see Fig. 2 for examples). Our experiments show that recent text-to-image generation models perform well at object recognition by generating high-fidelity objects but struggle at object counting and spatial relation understanding, with a large gap between the model performances and upper bound accuracy.

Second, we introduce social bias evaluation for text-to-image generation models. Recent work has reported social biases in vision-and-language datasets and models learned from them [50, 6]. We evaluate whether models trained on such datasets show bias when generating images from text. For this, we generate images of people with different professions that should not be related to a specific gender or skin tone (*e.g.*, nurse, doctor, teacher). Then, we detect gender, skin tone, and attributes from the generated images. We quantify biases by analyzing the distribution of the detected gender/skin tones and their relation to various professions/attributes. Our quantitative study shows that recent text-to-image models learned certain biases when generating images from some text prompts (*e.g.*, receptionist $\rightarrow$ female / plumber $\rightarrow$ male / female $\rightarrow$ wearing skirts / male $\rightarrow$ wearing suits). For automated gender and attribute detection, we use BLIP-2 [36] by asking visual questions (*e.g.*, "the person looks like a male or a female?"). For automated skin tone detection, we detect faces from images with FAN [8] and estimate illumination and facial albedo with TRUST [20]. Then we calculate Individual Typology Angle (ITA) [13] and find the closest skin tone in the MST scale [40]. Our final automated detection methods are highly correlated with human evaluation.

Our contributions can be summarized as follows: **(1)** We introduce PAINTSKILLS, a diagnostic evaluation dataset for text-to-image generation models, which allows carefully controlled measurement of the three fundamental visual reasoning skills. We show that recent models are relatively good at object recognition (generating a single object) skill, but a large gap exists between the performance of recent models and the upper bound accuracy in object counting and spatial relation understanding skills. **(2)** We introduce a gender and skin tone bias assessment based on automated and human evaluation. We show that recent models learn specific gender/skin tone biases from web image-text pairs.

Overall, our observations suggest that current text-to-image generation models are good initial contributions, but have several avenues for future improvements in learning challenging visual reasoning skills and understanding social biases. We hope that our evaluation work will allow the community to systemically measure such progress.

2. Related Works

Text-to-Image Generation Models. [38, 48] pioneered deep learning-based text-to-image generation. [48] introduced the GAN [22] framework to improve the visual reality of images. [71, 69] proposed to generate images in multiple stages by gradually increasing image resolution. Recently, the multimodal language model and diffusion model have been widely used for this task. X-LXMERT [14] and DALL-E [45] introduce multimodal transformer language models that learn the distribution of the sequence of discrete image codes given text input. Denoising diffusion models [54, 29, 49, 41] is another widely used model type in which a text-conditional denoising autoencoder iteratively updates noisy images into clean images. Recent multimodal language models (*e.g.*, Parti [70] and MUSE [12]) and diffusion models (*e.g.*, Stable Diffusion [49], DALL-E 2 [44], and Imagen [51]) deliver a high level of photorealism in a wide range of domains.

Metrics for Text-to-Image Generation. The text-to-image community has commonly used two types of automated evaluation metrics: image quality and image-text alignment. For image quality, Inception Score (IS) [52] and Fréchet Inception Distance (FID) [25] are the metrics most commonly used. They use the features of a pretrained image classifier such as Inception v3 [57] to measure the diversity and visual reality of the generated images. These metrics use a classifier pretrained on ImageNet [18] that mostly contains single-object images. Therefore, they are not suitable for more complex datasets [21]. To measure image-text alignment, metrics based on retrieval, captioning, and object detection models have been proposed. R-precision [69] evaluates the multimodal semantic relevance by the retrieval score of the original text given generated images with a pretrained image-to-text alignment model. [30, 26] employ an image caption generator to obtain captions for generated images and report language evaluation metrics such as BLEU [42] and CIDEr [61]. Semantic Object Accuracy (SOA) [26] measures whether an object detector can detect an object described in the text from a generated image. Evaluation based on R-precision and captioning can fail when different captions correctly describe the same image [26, 21].[2] In addition, unlike object detec-

[2] An image including 2 apples can be described as, "there are 2 apples"

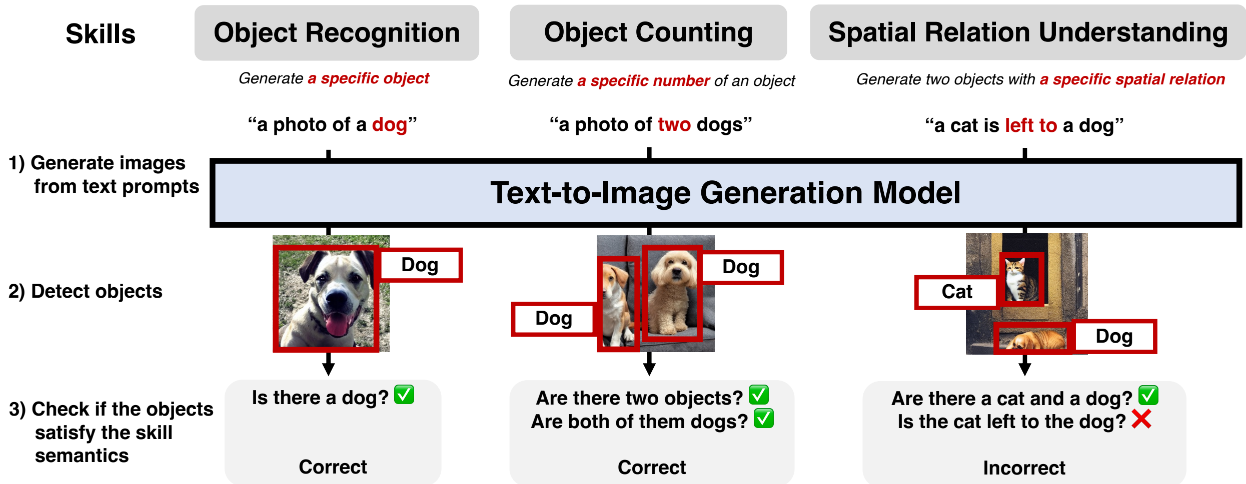


Figure 2. Illustration of the visual reasoning evaluation process with PAINTSKILLS (Sec. 3). We generate images from text prompts that require three different visual reasoning skills. Based on object detection results, we evaluate the visual reasoning capabilities of models by checking whether the generated images align with input text prompts. The example images are generated with Stable Diffusion.

tion, the retrieval/captioning models do not provide visually interpretable evidence of the scoring. SOA only focuses on the existence of objects, which makes it not well suited to evaluate object attributes and the relationship between objects [26, 21]. In contrast to existing alignment metrics, where reasoning based on alignment scoring is hard to understand, our PAINTSKILLS measures the text-to-image generation ability in a more fine-grained and transparent manner with three skills, including object recognition, object counting, and spatial relation understanding, to pinpoint model weaknesses.

Measuring Bias in Multimodal Models. While much research has been done on evaluating common social biases in image-only [65, 56] and text-only [74, 10] models, recent research work conduct such studies in multimodal models and datasets. [55, 50] showed social biases in visually grounded word embeddings. [6, 5, 58, 9, 73, 28, 62, 27] examine social biases in image-text datasets. [39] evaluate the diversity and inclusiveness of images containing people of specific occupations with respect to gender and race. [63, 68, 67, 6, 4] investigate biases in image-text retrieval models. Bansal *et al*. [3] and Zhang *et al*. [72] measure how text-to-image generation models behave differently with an intervention (*e.g*., adding phrases about gender, attributes, or skin color) to an original prompt. To our knowledge, our work provides the first evaluation metrics and analysis of measuring gender and skin tone biases in text-to-image generation models from diverse prompts with combinations of gender and professions, without prompt intervention.

or "two apples", which results in different values from text metrics.

3. PAINTSKILLS: A Diagnostic Evaluation Dataset for Compositional Visual Reasoning Skills

We introduce PAINTSKILLS, a diagnostic evaluation dataset for compositional visual reasoning skills of text-to-image generation models. Inspired by the recent vision-language skill-concept analysis of Whitehead *et al*. [66], we define three visual reasoning skills: object recognition, object counting, and spatial relation understanding.[3] To evaluate each skill, we calculate accuracy based on the detection results of the generated images, as illustrated in Fig. 2. In the following, we explain the skill definitions (Sec. 3.1) and the data collection process (Sec. 3.2).

3.1. Skills

Object Recognition. Given a text describing a specific object class (*e.g*., an airplane), a model generates an image that contains the intended class of object.

Object Counting. Given a text describing M objects of a specific class (*e.g*., 3 dogs), a model generates an image that contains M objects of that class.

Spatial Relation Understanding. Given a text describing two objects having a specific spatial relation (*e.g*., one is right to another), a model generates an image including two objects with the relation.

[3]There are other skills for image generation that the current three skills do not cover (*e.g*., text rendering). In this work, we focus on introducing skill-specific evaluation with object control skills fundamental to more complex skills.

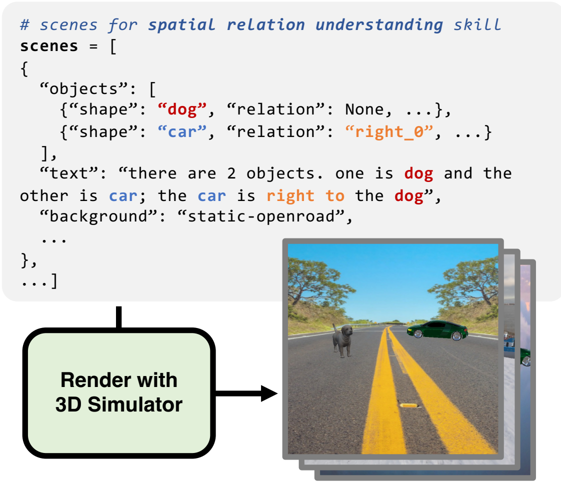


Figure 3. Dataset generation process (spatial relation understanding skill shown in this example) of PAINTSKILLS. For each skill, we generate scene configurations where object/attribute/layout combinations have a uniform distribution to avoid statistical shortcuts for reasoning. We use a 3D simulator for rendering images.

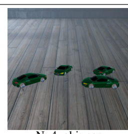
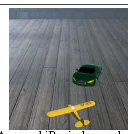
Skills	Object Recognition	Object Counting	Spatial Relation Understanding
Description	**a specific object**	**a specific number** of an object	two objects with **a specific spatial relation**
Template	`a photo of <obj>`	`a photo of <N> <obj>`	`a <objB> is <rel> a <objA>`
			
Keywords	obj: car	N: 4, obj: car	objA: car, objB: airplane, rel: below

Table 1. Example images, templates, and prompts of PAINTSKILLS. See appendix for more examples.

3.2. PAINTSKILLS Dataset Collection

The widely used visual question answering datasets such as VQA [2, 23] and GQA [31] are created by first collecting images, then collecting question-answer pairs from the images. However, since a few common objects dominantly appear in the image dataset, such data collection process results in a dataset with a highly skewed distribution towards a few common objects, questions, and answers. This often causes models trained on the datasets to depend on statistical bias instead of the desired compositional reasoning process [23, 1, 15, 17]. PAINTSKILLS addresses this problem by explicitly controlling the statistical bias between objects and input text. We collect text-image pairs for PAINTSKILLS in three steps: (1) We define scene configurations for each skill, in which the objects, attributes (*e.g.*, count), and relations are uniformly distributed. (2) We generate text prompts by creating templates with objects, numbers, and spatial relations. (3) We generate images from the scene configurations using a 3D simulator.

We develop the simulator using Unity[4] engine. The simulator takes a list of scene configurations and renders images from them. Each scene is represented as a list of objects, a text prompt, and a background, where each object has its own attributes, including class, location, and scale. Attributes can be specified or not. If an attribute is not specified, the simulator will use a default value or random sample from a uniform distribution while satisfying the other specified conditions. Backgrounds are sampled from 13 different images that do not contain object classes used in visual reasoning skill evaluation. We use 15 frequent object classes in MS COCO [37]: {`person, dog, airplane, bike, car, ...`}, object count range: {`1, 2, 3, 4`}, and 4 spatial relations: {`above, below, left, right`}.

As shown in Fig. 3, the simulator randomly assigns the object states (location, rotation, pose) and backgrounds, while satisfying the condition 'car is right to dog'. We generate 23,250/21,600/13,500 and 2,325/2,160/2,700 scenes for train and test splits of object recognition/object counting/spatial relation understanding skills, respectively. In Table 1, we provide sample images and corresponding text prompts for each skill in PAINTSKILLS. The text prompts are generated by composing keywords with a template.

Our simulator can be easily extended with custom objects and attributes. In the appendix, we provide the full prompt templates and detailed scene configurations including parameters, objects, and attributes.

4. Evaluations

We evaluate text-to-image generation models on two new criteria: visual reasoning skills (Sec. 4.1) and social biases (Sec. 4.2).

4.1. Visual Reasoning Skill Evaluation

As illustrated in Fig. 2, we evaluate models with three visual reasoning skills: object recognition (object), object counting (count), and spatial relation understanding (spatial). Following [26], we evaluate the skills based on how well an object detector can detect the object described in the input text. For each skill, we train a DETR [11] object detector. We initialize DETR parameters from the official checkpoint with ResNet101 [24] backbone trained on the MS COCO [37] *train 2017* split. In Table 2, we show the accuracy of DETR on the test split of each skill dataset, which is the upper bound performance. We also provide human evaluation results showing our proposed skill metrics align with human perception in Table 3.

Object Recognition. We evaluate the skill with average accuracy on N test images of whether an object detector correctly identifies the target class from the generated

[4] https://unity.com

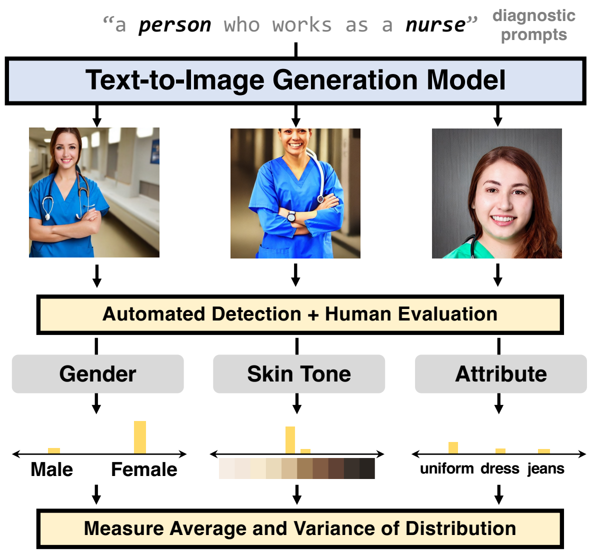


Figure 4. Overview of our social bias analysis (Sec. 4.2). Models generate images with a set of diagnostic prompts (*e.g.*, a person who works as a nurse), then with automated detectors and human evaluation, we estimate the gender, skin tone, and attributes shown in the images. Images in the examples were generated with Stable Diffusion.

images: $\frac{1}{N}\sum_i^N \mathbf{1}(o^{Det(i)} = o^{GT(i)}$ and $p^{Det(i)} > p^{th})$, where $o^{Det(i)}$ is a class that an object detection model predicts, $p^{Det(i)}$ is the classification confidence and $o^{GT(i)}$ is the ground-truth target object class.

Object Counting. We evaluate the skill with the average accuracy of whether an object detector correctly identifies the M objects of the target class from the generated images: $\frac{1}{N}\sum_i^N \mathbf{1}(o_j^{Det(i)} = o^{GT(i)}, \forall j \in \{1 \ldots M^{(i)}\})$, where $o_j^{Det(i)}$ is the class of the j-th object that an object detection model predicts, $o^{GT(i)}$ is target object class, and $M^{(i)}$ is the number of objects for the i-th image.

Spatial Relation Understanding. We evaluate the skill with the average accuracy of whether an object detector correctly identifies both target object classes and pairwise spatial relations between objects: $\frac{1}{N}\sum_i^N \mathbf{1}(o_1^{Det(i)} = o_1^{GT(i)}$ and $o_2^{Det(i)} = o_2^{GT(i)}$ and $rel^{Det(i)} = rel^{GT(i)})$, where $rel^{Det(i)}$ are the relation between two objects in the i-th image. We decide the spatial relation to be one of the four relations {`above`, `below`, `left`, `right`} based on the directions between two object positions from their 2D coordinates.

4.2. Social Bias Evaluation

As shown in Fig. 4, we measure the gender and skin tone biases of text-to-image generation models. For this, we first generate images from diagnostic prompts (Sec. 4.2.1), detect gender, skin tone, and attributes from the images (Sec. 4.2.2 and Sec. 4.2.3), and measure how they are skewed from an unbiased uniform distribution (Sec. 4.2.4).

4.2.1 Image Generation with Diagnostic Prompts

We create *diagnostic prompts* by composing a gender $G \in$ {`a man`, `a woman`, `a person`} and a profession $P \in$ {`accountant`, `engineer`, $\cdots$} (in total 83), using a template "G `who works as a/an` P". We also include three prompts without profession (just "G"), making 252 prompts (=3 × 83 + 3) in total; see appendix for the full list. The prompts starting with 'a man/woman' would reveal the bias of certain genders, and the prompts starting with 'a person' would reveal the bias of certain professions. We sample 9 images from a text-to-image generation model for each diagnostic prompt. From the generated images, we detect gender, skin tone, and attributes using automated detection models and verify the reliability of detection models with human evaluation (see appendix).

4.2.2 Detection Categories

Gender. For gender bias analysis, we use two *gender* categories: {`male`, `female`}. A wide range of genders is beyond the scope of finite categories [32]. However, even humans cannot reliably estimate the gender of other people across a wide spectrum of gender categories based only on appearance. Hence, following concurrent work [72, 3], we limit our gender categorization to binary for the current study, where we focus on exposing different types of bias in text-to-image generation models.

Skin Tone. Next, our skin tone analysis uses the Monk Skin Tone (MST) Scale [40], which transforms the continuous skin tone spectrum into 10 tones. Such fine-grained skin tone scales can better reflect a diversity of communities than binary categorizations such as 'light' and 'dark' skin. Although one may categorize people into racial categories (*e.g.*, Black, White, *etc.*), race is not a biological concept and should be understood as a socially constructed and political concept [16, 7]. Because race is not naturally inherent, fixed, or mutually exclusive [7, 46], inferring one's racial identity from appearance and assuming that one's race falls into a single category could lead to an inaccurate inference of one's racial identity.

Attribute. Lastly, we analyze the 15 attributes from Zhang *et al.* [72]. We use the frequency of the attributes detected to measure the difference in the presentation of different genders, skin tones, and professions.

4.2.3 Automated Detection and Human Evaluation

We detect gender, skin tone, and attributes from the generated images using automated detection models and verify their reliability with human evaluation. We experiment with different detection models for gender, skin tone, and attributes to compare their accuracy and reliability. The following describes how we use the finally chosen detection models. See appendix for a detailed comparison between models and human evaluation.

Gender Detection. We use BLIP-2 [36] to detect gender in the generated images, by asking the question `"the person looks like a male or a female?"`[5] and then detect whether BLIP-2 returns male/female in the answer. In our experiments, BLIP-2 showed less bias and higher accuracy than CLIP (ViT/B-32) [43] in COCO bias testing [63] and Adience gender dataset [19] (82% BLIP-2 *vs*. 66% CLIP; see appendix for more details).

Skin Tone Detection. We use FAN [8] to detect facial landmarks in the generated images, and use TRUST (BalancedAlb checkpoint) [20] to estimate the illumination of the images and albedo UV map of the facial crops. We take illumination into account when detecting skin tone, as raw pixel values are a function of both the scene lighting and the subject's true skin tone [53]. On the detected facial albedo UV maps, we calculate the Individual Typology Angle (ITA) [13] based on L* (lightness) and B* (yellow/blue) components of the CIE-L*a*b* colorspace and find the closest skin tone in MST scale (1-10) [40]. In our experiments, using facial landmarks and addressing illumination improves the accuracy of skin tone detection (see appendix for more details).

Attribute Detection. We give BLIP-2 an image and a question, `"Is the person wearing` A`?"` for each attribute A (e.g. `"a suit"`, `"jeans"`) and check if the model responds with "yes". In our experiments, BLIP-2 is more accurate than CLIP-based classification [72] in attribute detection (92% BLIP-2 *vs*. 79% CLIP; see appendix for details).

4.2.4 Measuring Bias: Average and Variance

From the detection results, we obtain distributions for gender (binary), skin tone (10-way categorical), and attribute (binary for each item). To show to which gender, skin tone, and attribute category the distribution is skewed, we report the average value of each bias category. To compute the overall bias distribution, we use mean absolute deviation (MAD) that measures the distance between detected gender/skin tone category/attribute distributions and unbiased uniform distribution: $\frac{1}{N}\sum_{i=1}^{N}|p_i - \bar{p}|$, where $p_i \in [0,1]$ are the normalized counts of the i-th gender or skin tone category, $\bar{p}$ is the mean normalized counts (0.5 for gender; 0.1 for skin tone), and N is the number of gender/skin tone scales (2 for gender; 10 for skin tone). MAD is minimized to 0 when the category distribution is uniform (unbiased) and maximized when the category distribution is one-hot (entirely biased to a single category).

[5] We experimented with several prompts and found this produces the best results.

Evaluator	Images	Skill Accuracy (%) (↑)			
		Object	Count	Spatial	Avg.
DETR	GT (oracle)	100.0	97.8	96.2	98.0
	GT shuffled (random)	6.3	1.7	0.3	2.8
	DALL-E$^{\text{Small}}$	57.5	18.2	2.4	26.0
	minDALL-E	89.9	**47.5**	**50.7**	**62.7**
	Stable Diffusion	**96.2**	37.8	7.9	47.3

Table 2. DETR evaluation on images generated from the T2I models finetuned on PAINTSKILLS.

	Evaluator	Images	Skill Accuracy (%) (↑)			
			Object	Count	Spatial	Avg.
(A)	Human	DALL-E$^{\text{Small}}$	52.0	42.0	4.0	30.7
		minDALL-E	86.0	**64.0**	**64.0**	**68.7**
		Stable Diffusion	**94.0**	48.0	16.0	54.7
(B)	DETR	DALL-E$^{\text{Small}}$	64.0	34.0	0.0	28.0
		minDALL-E	86.0	**54.0**	**66.0**	**64.0**
		Stable Diffusion	**98.0**	44.0	4.0	54.0

Table 3. Human and DETR evaluation on PAINTSKILLS. For each skill, we sample 50 images, collecting 3x50 = 150 images for each model.

5. Experiments and Results

We introduce the evaluated text-to-image generation models in Sec. 5.1, then show the evaluation results of visual reasoning skills (Sec. 5.2) and social biases (Sec. 5.3).

5.1. Evaluated Models

Since the pretrained checkpoints of the original DALL-E model have not been released at the time of this analysis, we experiment with two different publicly available implementations of DALL-E: DALL-E$^{\text{Small}}$ [64] and minDALL-E [33]. The models consist of a discrete VAE (dVAE) [34, 59, 47] that encodes images with grids of discrete tokens and a multimodal transformer that learns the joint distribution of text and image tokens. We also experiment with Stable Diffusion v1.4 [49] and Karlo [35], recent state-of-the-art diffusion models that publicly released their checkpoints. As Karlo has not released its training code, we use it only for social bias evaluation. We provide more details about each model in the appendix.

5.2. Visual Reasoning Skill Results

Object Detector Accuracy. In the top rows of the Table 2, we show the visual reasoning accuracy on the ground-truth

Skills	Object Recognition		Object Counting		Spatial Relation Understanding	
Prompts	‘a **dog**’	‘a **bicycle**’	‘**3** dogs’	‘**2** bicycles’	‘a suitcase is **left to** a person’	‘an umbrella is **right to** a stop sign’
GT						
DALL-E$^{\text{Small}}$						
minDALL-E						
Stable Diffusion						

Table 4. Images generated by three text-to-image generation models finetuned on PAINTSKILLS. Objects detected from the images are shown in colored bounding boxes.

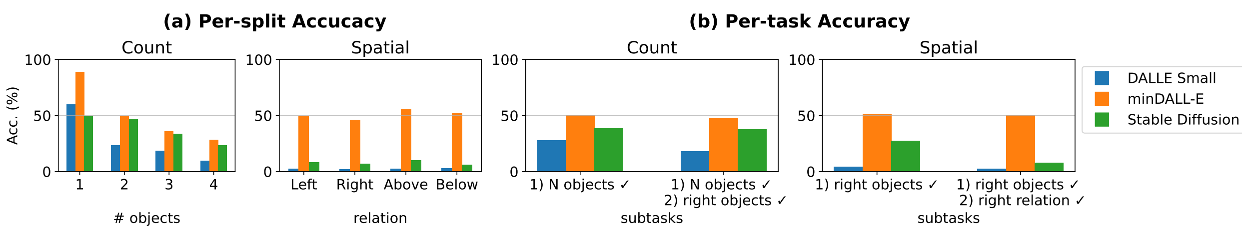


Figure 5. Detailed analysis of *count* and *spatial* skills of 3 models, in terms of (a) per-split and (b) per-task accuracy.

(GT) PAINTSKILLS images and randomly shuffled GT images. With a high average oracle accuracy of 98.0%, we expect our evaluation to serve as good automated metrics for visual reasoning skills. The low average accuracy of randomly shuffled GT images (2.8%) indicates that a model cannot achieve a high score on PAINTSKILLS without correct placement of objects.

Which model is good at which skill? Table 2 shows that Stable Diffusion achieves the highest accuracy of 96.2% in *object* skill. This could be explained by its high-fidelity image generation based on the largest training data (5B) and highest resolution (512x512). However, in *count* and *spatial* skills, minDALL-E achieves better accuracy than Stable Diffusion. As shown in Table 4, even though Stable Diffusion could generate high-fidelity objects, the model often generates more (5 instead of 3 dogs) or fewer (1 instead of 2 bicycles) objects than the number described in the prompt. Likewise, Stable Diffusion often misses an object (person, umbrella) described in prompts for *spatial* skill. Overall, a huge gap exists between the performance of all models and the upper bound accuracy on *count*/*spatial* skills, indicating a large room for improvement.

Fine-grained Skill Analysis. Fig. 5 (a) shows the per-split accuracy of *count* and *spatial* skills. In *count* skill, the models score lower accuracy with prompts with more objects. In *spatial* skill, the models achieve similar accuracy for all four spatial relations. Fig. 5 (b) shows the per-task accuracy of the two skills. In *count* skill, a model needs to 1) generate the correct number of objects and 2) ensure all objects are in the right classes. For all three models, the accuracy difference between 1) and 1) + 2) is small, indicating that the bottleneck for this task is 1) generating the right number of objects rather than 2) generating the correct objects. In *spatial* skill, a model needs to 1) generate two right objects of the right classes and 2) satisfy the given spatial relation. Stable Diffusion shows a larger drop between 1) and 1) +

Figure 6. Gender, skin tone, and attribute detection results with automated and expert human evaluation. The images are generated by the Stable Diffusion model, using the gender/skin tone-neutral prompts (*e.g.*, "a person who works as a biologist"). For gender estimation, both automated detection and human evaluation agreed on all examples here. For attribute and skin tone estimation, automated detection and human annotation are closely aligned in most cases. The detection results are presented in order of top-left → top-right → bottom-left → bottom-right. *M: Male, F: Female, Y: Yes, N: No.*

Training data	Model	Skill Accuracy (%) (↑)			
		Object	Count	Spatial	Avg.
100%	minDALL-E	89.9	47.5	50.7	62.7
	Stable Diffusion	96.2	37.8	7.9	47.3
50%	minDALL-E	90.1	49.4	53.3	64.3
	Stable Diffusion	96.0	42.2	7.6	48.6
10%	minDALL-E	90.8	50.9	38.2	60.0
	Stable Diffusion	94.2	37.9	8.9	47.0

Table 5. PAINTSKILLS DETR-based accuracy of minDALL-E and Stable Diffusion v1.4 with different scales of training data.

2) accuracy, indicating that differentiating the four spatial relations is the bottleneck for this model.

Human Evaluation. To verify if our DETR-based evaluation aligns with human perception, we ask a human expert to evaluate the images generated from the models finetuned on PAINTSKILLS. The expert evaluated 150 images for each skill (3 models x 50 images). In Table 3, we find that DETR-based evaluation achieves similar accuracy with the human evaluation in all three models, and relative performance between models is the same in both evaluations.

Does PAINTSKILLS have enough finetuning data? As evaluation with PAINTSKILLS involves finetuning, we experiment with finetuning with different numbers of training data to see whether text-to-image generation models see enough training examples to learn skills and avoid domain gaps (*e.g.*, real *vs*. synthetic images). Table 5 shows that model performances between 100% and 50% of the data are similar, indicating that PAINTSKILLS training dataset is large enough for the models to adapt.

5.3. Social Bias Results

As described in Sec. 4.2 and Fig. 4, we generate images with text-to-image generation models[6] from diagnostic prompts (*e.g.*, "a person who works as a nurse"). In Fig. 6, we show examples of gender, skin tone, and attribute detection based on automated methods and human annotators. Please see appendix for our human evaluation of the accuracy and reliability of automated detectors.

Gender Bias. Table 6 shows the per-profession and average gender bias of three models. While all three models have an overall tendency to generate male images, models have different gender biases in different professions. For example, from 'Singer' prompts, minDALL-E tends to generate more male images, whereas and Karlo and Stable Diffusion tend to generate more female images.

The 'gender' column of Table 8 column shows that minDALL-E achieves lower MAD than Karlo and Stable Diffusion, indicating that Karlo and Stable Diffusion have a stronger tendency to generate images of a specific gender from gender-neutral prompts than minDALL-E.

Table 9 compares the attribute presence for gender prompts. All three models tend to generate skirts only for woman prompts, and tend to generate suit/jacket/tie more frequently for man prompts.

Skin Tone Bias. Table 7 shows three models' per-profession/average skin tone bias. Unlike the gender bias

[6]For social bias analysis, we only experiment with images from minDALL-E, Stable Diffusion, and Karlo, because we find that the visual quality of images from DALL-E$^{\text{Small}}$ is highly distorted and does not provide meaningful semantics.

Profession	Average Gender (male: -1 / female: +1)		
	minDALL-E	Karlo	Stable Diffusion
Engineer	-0.78	-1.0	-1.0
Library assistant	-0.11	1.0	1.0
Scientist	-0.11	0.56	-0.33
Singer	-0.33	0.33	0.56
Baker	-0.11	-0.33	0.33
Average	-0.25	-0.22	-0.42

Table 6. Per-profession examples and average gender bias of images generated from gender-neutral prompts: 'a person who works as a/an [profession]'. -1 and 1 refer to male and female, respectively. See appendix for the full table.

Profession	Average Skin Tone (1-10)		
	minDALL-E	Karlo	Stable Diffusion
Judge	5.13	5.05	5.04
Miner	5.5	5.18	5.59
Porter	5.33	5.55	5.44
Secretary	5.05	5.0	5.0
Tailor	5.09	5.44	5.31
Average	5.19	5.13	5.14

Table 7. Per-profession examples and average skin tone bias of images generated from prompts: 'a [person/man/woman] who works as a/an [profession]'. We use Monk Skin Tone (MST) Scale of 1-10 [40]. See appendix for the full table.

results in Table 6, where different professions correlate differently with genders, all three models tend to generate images with similar skin tones for all professions. All models generate tones around 5 and 6, indicating very light and dark skin tones are marginalized from the learned representation of the models. See appendix for the skin tone analysis per attributes.

The 'skin tone' column of Table 8 shows that all three models achieve similar MAD, while minDALL-E achieves the lowest value. The MAD of *N-hot* distributions of 10-category of are as follows: $\text{MAD(1-hot)} = 0.18, \text{MAD(2-hot)} = 0.16, \text{MAD(3-hot)} = 0.14, \cdots, \text{MAD(10-hot=uniform)} = 0$. As the models show MAD between 0.16 and 0.18, their skin tone distributions are similar to 1-hot and 2-hot distributions with a concentration on the MST scales of 5 and 6.

6. Conclusion

We propose two new evaluation aspects of text-to-image generation: visual reasoning skills and social biases. For visual reasoning skills, we introduce PAINTSKILLS, a compositional diagnostic evaluation dataset designed to measure three skills: object recognition, object counting, and spatial relation understanding. Our experiments show that recent text-to-image models perform better in recognizing objects than object counting and understanding spatial relations, while a large gap exists between the model performances and upper bound accuracy in the latter two skills. We also show that the models have learned specific gender/skin tone biases from web image-text pairs. We hope our evaluation provides novel insights for future research on learning challenging visual reasoning skills and understanding social biases.

Model	MAD (↓)	
	Gender	Skin Tone
uniform (unbiased)	0.0000	0.0000
minDALL-E	**0.1984**	**0.1687**
Karlo	0.3545	0.1707
Stable Diffusion	0.3618	0.1698
one-hot (entirely biased)	0.5000	0.1800

Table 8. Comparison of overall gender and skin tone bias of each model. MAD measures the distance between detected gender/skin tone distribution and an unbiased uniform distribution. The best (lowest) values are bolded.

Model	Prompts	Attributes (presence: 1 / absence: 0)			
		skirt	suit	jacket	tie
minDALL-E	Woman	0.1	0.12	0.11	0.02
	Man	0.0	0.39	0.29	0.23
	Woman - Man	+0.1	-0.27	-0.18	-0.21
Karlo	Woman	0.05	0.16	0.02	0.0
	Man	0.0	0.27	0.17	0.18
	Woman - Man	+0.05	-0.11	-0.15	-0.18
Stable Diffusion	Woman	0.07	0.19	0.07	0.0
	Man	0.0	0.35	0.26	0.2
	Woman - Man	+0.07	-0.16	-0.19	-0.2

Table 9. Presence of attributes for images from gender-specific prompts: 'a [man/woman] who works as a/an [profession]'. The 'Woman - Man' rows show the relative differences in attribute presence between two gender-specific prompts (*i.e.* negative/positive values indicate the attributes are more correlated to woman/man, respectively). See appendix for more attributes.

Acknowledgments

We thank Heesoo Jang, Peter Hase, Hyounghun Kim, Adyasha Maharana, and Yi-Lin Sung for their helpful comments. This work was supported by ARO Award W911NF2110220, DARPA MCS Grant N66001-19-2-4031, ONR Grant N00014-23-1-2356, and a Google Focused Research Award. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *CVPR*, 2018. 2, 4

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 4

[3] Hritik Bansal, Da Yin, and Masoud Monajatipoor. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? In *EMNLP*, 2022. 3, 5

[4] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only, Nov. 2022. Association for Computational Linguistics. 3

[5] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *ArXiv*, abs/1912.00578, 2019. 3

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv*, abs/2110.01963, 2021. 2, 3

[7] Simone Browne. *Dark Matters: On the Surveillance of Blackness*. Duke University Press, 2015. 5

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 2, 6

[9] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. *ArXiv*, abs/1803.09797, 2018. 3

[10] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 3

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *ECCV*, 2020. 2, 4

[12] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-To-Image Generation via Masked Generative Transformers. pages 1–22, 2023. 2

[13] A. CHARDON, I. CRETOIS, and C. HOURSEAU. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13(4):191–208, 1991. 2, 6

[14] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *EMNLP*, 2020. 2

[15] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *EMNLP*, 2019. 2, 4

[16] Kate Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021. 5

[17] Corentin Dancette, Remi Cadene, Xinlei Chen, and Matthieu Cord. Overcoming Statistical Shortcuts for Open-ended Visual Counting. *ArXiv*, arXiv:2006.10079v2, 2020. 2, 4

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[19] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014. 6

[20] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *ECCV*, 2022. 2, 6

[21] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. Adversarial Text-to-Image Synthesis: A Review. *Neural Networks*, 144:187–209, jan 2021. 1, 2, 3

[22] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *NIPS*, 2014. 2

[23] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, 2017. 2, 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 4

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*, 2017. 1, 2

[26] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic Object Accuracy for Generative Text-to-Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 2, 3, 4

[27] Yusuke Hirota, Yuta Nakashima, and Noa García. Gender and racial bias in visual question answering datasets. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. 3

[28] Y. Hirota, Y. Nakashima, and N. Garcia. Quantifying societal bias amplification in image captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13440–13449, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 2

[30] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. In *CVPR*, 2018. 1, 2

[31] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 4

[32] Os Keyes, Chandler May, and Annabelle Carrell. You keep using that word: Ways of thinking about gender in computing research. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. 5
[33] Saehoon Kim, Sanghun Cho, Chiheon Kim, Doyup Lee, and Woonhyuk Baek. mindall-e on conceptual captions. https://github.com/kakaobrain/minDALL-E, 2021. 1, 6
[34] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *NIPS*, 2013. 6
[35] Donghoon Lee, Jiseob Kim, Jisu Choi, Jongmin Kim, Minwoo Byeon, Woonhyuk Baek, and Saehoon Kim. Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo, 2022. 1, 6
[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 2, 6
[37] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 4
[38] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating Images from Captions with Attention. In *ICLR*, 2016. 2
[39] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. *Diversity and Inclusion Metrics in Subset Selection*, page 117–123. Association for Computing Machinery, New York, NY, USA, 2020. 3
[40] Ellis Monk. Monk Skin Tone Scale. https://skintone.google, 2022. 2, 5, 6, 9
[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 2022. 2
[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Weijing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 2002. 2
[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Jong Wook, Kim Chris, Hallacy Aditya, Ramesh Gabriel, Goh Sandhini, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 6
[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, 2204.06125, 2022. 2
[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *ICML*, 2021. 1, 2
[46] Victor Ray. *On Critical Race Theory: Why It Matters & Why You Should Care*. Random House Publishing Group, 2022. 5
[47] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *NeurIPS*, 2019. 6
[48] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2
[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, June 2022. 1, 2, 6
[50] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *NAACL*, 2021. 2, 3
[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*, 2022. 2
[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *NIPS*, 2016. 1, 2
[53] Candice Schumann, Gbolahan O Olanubi, Auriel Wright, Ellis Monk Jr, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ml fairness. *arXiv preprint arXiv:2305.09073*, 2023. 6
[54] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2
[55] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *ArXiv*, abs/2104.08666, 2021. 3
[56] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 701–713, New York, NY, USA, 2021. Association for Computing Machinery. 3
[57] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016. 2
[58] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, WWW '21, page 633–645, New York, NY, USA, 2021. Association for Computing Machinery. 3
[59] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *NIPS*, 2017. 6
[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS*, 2017. 1
[61] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, nov 2015. 2

[62] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 324–335, 2022. 3

[63] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *EMNLP*, 2021. 3, 6

[64] Phil Wang. Dalle-pytorch. https://github.com/lucidrains/DALLE-pytorch, 2021. 1, 6

[65] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, pages 5309–5318, 2019. 3

[66] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating Skills and Concepts for Novel Visual Question Answering. In *CVPR*, 2021. 3

[67] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. 3

[68] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. 3

[69] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *CVPR*, 2018. 1, 2

[70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research*, 2022. 2

[71] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN : Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In *ICCV*, 2017. 2

[72] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models, 2023. 3, 5, 6

[73] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[74] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, pages 2979–2989, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 3