HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models

Eslam Mohamed Bakr[1], Pengzhan Sun[2] *, Xiaoqian Shen [1] *,
Faizan Farooq Khan [1] *, Li Erran Li[3], Mohamed Elhoseiny[1]
{eslam.abdelrahman, xiaoqian.shen, faizan.khan,
mohamed.elhoseiny}@kaust.edu.sa
pengzhan@comp.nus.edu.sg, lilimam@amazon.com
[1]King Abdullah University of Science and Technology (KAUST)
[2]National University of Singapore [3]AWS AI, Amazon

Abstract

*In recent years, Text-to-Image (T2I) models have been extensively studied, especially with the emergence of diffusion models that achieve state-of-the-art results on T2I synthesis tasks. However, existing benchmarks heavily rely on subjective human evaluation, limiting their ability to holistically assess the model's capabilities. Furthermore, there is a significant gap between efforts in developing new T2I architectures and those in evaluation. To address this, we introduce HRS-Bench, a concrete evaluation benchmark for T2I models that is **H**olistic, **R**eliable, and **S**calable. Unlike existing benchmarks that focus on limited aspects, HRS-Bench measures 13 skills that can be categorized into five major categories: accuracy, robustness, generalization, fairness, and bias. In addition, HRS-Bench covers 50 scenarios, including fashion, animals, transportation, food, and clothes. We evaluate nine recent large-scale T2I models using metrics that cover a wide range of skills. A human evaluation aligned with 95% of our evaluations on average was conducted to probe the effectiveness of HRS-Bench. Our experiments demonstrate that existing models often struggle to generate images with the desired count of objects, visual text, or grounded emotions. We hope that our benchmark help ease future text-to-image generation research. The code and data are available at https://eslambakr.github.io/hrsbench.github.io/.*

1. Introduction

Text-to-Image Synthesis (T2I), one of the essential multimodal tasks, witnessed remarkable progress starting from conditional GANs [58, 44, 74, 29, 78], which are shown to work on simple datasets [46, 69, 72, 36], to recently diffusion models [20, 61, 18, 10, 57, 55, 77, 45, 59], which are trained on large-scale datasets, e.g., LAION [63, 62].

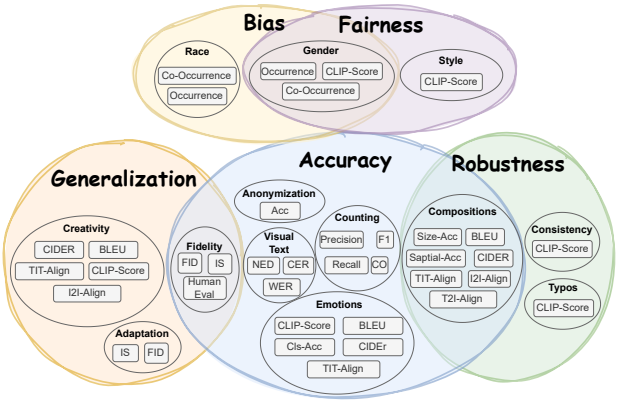

Figure 1: An overview of our proposed benchmark, HRS-Bench, measures 13 skills which could be grouped into five major categories; accuracy, robustness, generalization, fairness, and bias.

Despite the rapid progress, the existing models face several challenges, e.g., they cannot generate complex scenes with the desired objects and relationship composition [25, 37]. Furthermore, assessing the T2I models should include more than just fidelity, e.g., the ability to compose multiple objects and generate emotionally grounded or creative images. Therefore, some recent efforts are focusing on improving the existing metrics [23] or proposing new metrics that cover new aspects, such as bias [75], compositions [37, 25, 49]. Moreover, some other works propose new benchmarks, summarized in Table 1, that assess different aspects, e.g., counting [61, 12, 51], social-bias [12], and object fidelity [23, 51]. Even with various benchmarks available, they tend to only cover a limited range of aspects while overlooking crucial evaluation criteria such as robustness,

*Equal Contribution

Table 1: Comparison of text-to-image benchmarks in terms of: 1) The number of evaluated models. 2) The number of the covered skills. 3) The number of utilized metrics. 4) The evaluation type, whether human or metric based or both. 5) The number of prompts. 6) The prompt generation type, whether a template, human-based or both. 7) Whether there are different hardness levels are included. 8) Number of annotators contributed to the evaluation.

Method	# Models	# Skills	# Metrics	Eval Type		# Prompts	Prompt Type		Hardness Levels	# Annotators
				Human	Auto		Template	Human		
DrawBench [61]	5	4	0	✓	✗	200	✗	✓	✗	25
DALL-EVAL [12]	4	5	3	✓	✗	7330	✓	✗	✗	6
HE-T2I [51]	2	3	0	✓	✗	90	✗	✓	✓	20
TISE [23]	7	3	5	✗	✓	N/A	✗	✓	✗	N/A
HRS-Bench (Ours)	**9**	**13**	**17**	✓	✓	**45000**	✓	✓	✓	**1000**

fairness, and creativity.

To bridge this gap, we propose our Holistic, Reliable, and Scalable benchmark, dubbed HRS-Bench. In contrast to existing benchmarks, we measure a wide range of different generative capabilities, precisely 13 skills which can be grouped into five major categories, as demonstrated in Figure 1; accuracy, robustness, generalization, fairness, and bias. Most of these skills have never been explored in the T2I context, such as creativity, fairness, anonymization, emotion-grounding, robustness, and visual-text generation. Even the other previously explored skills were studied from a limited perspective, for instance, DALL-EVAL [12] studied the social bias by generating limited template-based prompts; only 145 prompts. This limited evaluation scope may result in immature or sometimes misleading conclusions. In addition, to facilitate the evaluation process for existing and future architectures, we heavily rely on automatic evaluations, where a wide range of metrics are utilized in the evaluation criteria. Moreover, HRS-Bench covers 50 scenarios, e.g., fashion, animals, transportation, and food. Figure 2 demonstrates the top 15 applications and their object distribution. We evaluate nine recent large-scale T2I models, i.e., Stable-Diffusion V1 [59] and V2 [3], DALL-E 2 [55], GLIDE [45], CogView-V2 [21], Paella [57], minDALL-E [2], DALL-E-Mini [1], and Struct-Diff [25]. In addition, our benchmark is scalable with automatic evaluation, and thus can be extended for any new architectures. To probe the effectiveness of our HRS-Bench, we conduct a human assessment that aligns well with our evaluations by 95% on average. Our contributions can be summarized as follows:

- We develop a Holistic, Reliable, and scalable T2I benchmark called HRS-Bench, depicted in Figure 1, which assess 13 skills covering 50 scenarios.
- Propose a new T2I alignment metric, called AC-T2I, which overcomes the composition limitations of existing large Vision-Language Models (VLMs) [31, 73].
- Nine T2I models are assessed based on our benchmark, including commercial and open-sourced ones.
- We verify the effectiveness of our HRS-Bench metric by conducting a human evaluation for 10% of our data per skill that shows excellent alignment.
- Driven by these holistic evaluations, several conclusions and findings are discussed. For instance, existing models often struggle to generate images with the desired count of objects, visual text, or grounded emotions.

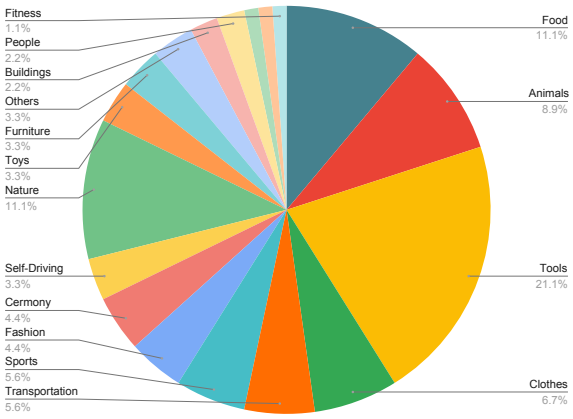

Figure 2: Pie chart demonstrates the wide range of covered scenarios by our proposed benchmark, termed HRS-Bench, and their object distribution.

2. Revisiting Text-to-Image Benchmarks

Recently, there have been rapid efforts toward designing better image-generation models [20, 61, 18, 10, 57, 55, 77, 45, 59]. However, most existing metrics suffer from several limitations that make them unreliable [23, 7, 22, 6, 75, 49, 66, 73]. Table 1 summarizes the existing T2I benchmarks.
–**DrawBench.** Imagen [61] proposes DrawBench to evaluate the T2I models from other aspects along with the image quality. Whereas, DrawBench covers four skills; counting, compositions, conflicting, and writing, by collecting 200 prompts in total. Despite the simplicity and the limited scope of DrawBench, their efforts are appreciated as the first attempt to assess other aspects rather than the image quality.
–**DALL-EVAL.** **[12]** It proposes a toolkit called PAINTSKILLS, which assesses three simple visual

reasoning skills; object recognition, object counting, and spatial relation understanding, alongside two social bias skills; gender and racial bias. To facilitate the automatic evaluation, they built a unity simulator including limited objects to collect the datasets, approximately 80 object classes. In contrast, we cover more than 700 object classes. DETR [9] is utilized for visual reasoning skills evaluation after being fine-tuned on the synthetic dataset collected from the simulator. However, DALL-EVAL evaluates only four models using only 7k prompts and six annotators. Therefore, there is a need for a more comprehensive benchmark that can take into account a broader range of models, prompts, and annotators to provide a more thorough evaluation.

–**HE-T2I. [51]** It proposes 32 possible aspects to benchmark T2I models. However, only three are evaluated; counting, shapes, and faces, and the rest are left unexplored. The three aspects are evaluated using human evaluations, where twenty annotators have contributed to the evaluation.

–**TISE. [23]** It introduces a bag of metrics to evaluate T2I models from three aspects; positional alignment, counting, and fidelity. In addition, three fidelity metrics are introduced, i.e., IS^*, OIS, and $OFID$, and two alignment metrics PA for positional alignment, and CA for counting alignment.

3. HRS-Bench

In this section, we first dissect the skills definition, then demonstrate the prompts collection pipeline.

3.1. Skills

3.1.1 Accuracy

–**Counting.** A reliable T2I model should be able to ground the details in the prompt. One form of these details is objects binding with a specific frequency, e.g., "four cars are parked around five benches in a park".

–**Visual Text.** Another essential aspect of assessing the model is generating high-quality text in wild scenes, e.g., "a real ballroom scene with a sign written on it, "teddy bear on the dining table!"". The importance of such skill comes from intervening in many scenarios, e.g., education applications, preparing illustration content, and designing billboards.

–**Emotion.** We measure to what extent the model can generate emotion-grounded images [5, 4, 41, 40], e.g., "a rainy scene about cake, which makes us feel excitement."

–**Fidelity.** Image fidelity indicates how accurately an image represents the underlying source distribution [64].

3.1.2 Robustness

To assess the T2I model's robustness, we cover two types of transformations; invariance and equivariance.

Invariance. Two skills are introduced to measure the invariance robustness: consistency and typos.

–*Consistency.* We measure the sensitivity of different T2I models towards prompt variations while keeping the same meaning and semantics, i.e., paraphrasing. For instance, generated images from these prompts, "a woman is standing in front of a mirror, carefully selecting the perfect handbag" and "in front of a mirror, a woman is selecting the perfect handbag for the day" should hold the same semantics.

–*Typos.* Two natural perturbations are utilized to assess the models against possible sensibleness noise that users could cause during inference, i.e., typos and wrong capitalization.

Equivariance. Three different compositions are explored for the equivariance robustness. Specifically, we study three types of compositions, i.e., spatial, attribute-specific, and action compositions.

–*Spatial composition.* In contrast to the counting skill which only measures the models' ability to compose multiple objects into a coherent scene, spatial composition additionally measures their ability to ground the detailed spatial relationship instructions mentioned in the input prompt, e.g., "a person and a dog in the middle of a cat and a chair".

–*Attribute-specific composition.* Two types of attributes are controlled to study the attribute binding ability, i.e., colors and size attributes. For instance, "an orange cat, a red dog, and a blue chair" and "a banana which is smaller than a person and bigger than a car", for colors and size attribute binding, respectively.

–*Action composition.* It incorporates different subjects that doing different actions, e.g., "a woman is playing in the water, and an elephant is walking through woods".

3.1.3 Generalization

–**Creativity.** In this skill, models aim to generate images that represent not only the textual description but are also imaginative and novel. The creativity skill can be regarded as an out-of-distribution generation [30]. Accordingly, we devised innovative text prompts that are conceptually plausible but may need to be more readily available in standard training data sources, detailed later in Section 3.2.

3.1.4 Fairness

We define fairness as the performance disparity among different sub-groups [24, 52]. A fair model should achieve the same performance on an arbitrary metric since no correlation exists between the metric and the protected attribute. Two attributes have been studied, i.e., gender and style. Following [12], gender refers to sex [32, 54] not the gender identity [43, 17]. We use two gender categories; male and female. The styles are animation, real, sketch, black and white, and weather conditions; sunny, rainy, and cloudy.

3.1.5 Bias

We assess the spurious correlation of the model towards pre-defined attributes, i.e., gender, race, and age. Using agnostic prompts towards a specific attribute, e.g., gender, the model should produce balanced generations of different classes of this attribute. For instance, the gender agnostic prompt could be, "Two persons are working on their laptops".

3.2. Prompts Collection

For each skill, we collect 3k prompts. To ensure our benchmark is holistic enough, we split the prompts equally into three hardness levels, i.e., easy, medium, and hard. In addition, half of the prompts are human-based, and the other half is template-based, depicted in Figure 3. We filter human prompts manually from existing datasets [70, 33, 31]. We use the foundation model GPT-3.5 [47], text-davinci-003* to facilitate prompts generation, which will be abbreviated as GPT-3.5 later for convenience.

–**Fidelity.** The human prompts are sifted from [70]. Whereas, the template-based prompts are created by defining a template that describes a styled scene that contains some objects, as shown in Figure 3. Then, we create the meta-prompt by sampling the styles from pre-defined styles and the objects from LVIS dataset [26]. Finally, GPT-3.5 [47] is utilized to generate the final prompts.

–**Consistency and Typos.** Consequently, the fidelity prompts are fed to Parrot [14] and NLU augmenter [19] to produce augmented prompts for consistency and typos, respectively. For consistency, we differentiate between the three hardness levels based on the similarity between the fidelity prompt and the augmented prompts using RoBERTa [38]. For typos, the number of introduced typos controls the three hardness levels, i.e., 1-2, 3-4, and 5-6, respectively.

–**Counting.** Given a meta-prompt, GPT-3.5 generate a realistic scenario that contains N objects. To generate the meta-prompts, we randomly samples the number of objects and the objects classes from LVIS dataset [26].

–**Visual Text.** We utilize GPT-3.5 to generate short descriptions which fit on a sign in a crowded scene. Then, we control the hardness levels by the text length and the surrounding scene complexity.

–**Emotion.** We sample random objects from LVIS [26], then append an emotion indicator word forming the meta-prompt. Finally, GPT-3.5 is utilized to generate the final prompts.

–**Creativity.** We craft text prompts that are challenging yet still within the realm of imagination. For easy level, we obtain subject, object, and relationship from Visual Genome [33] and obtain triplets by different combinations. Then sift through all the combinations from triplets extracted from LAION [62] dataset, and retain only the uncommon triplets. For the medium level, we fed the uncommon triplet to GPT-3.5 with the instruction: "Describe `subject`, `relation` and `object` in an imaginative way that will never be seen in the real world" and manually filter the undesirable sentences. Finally, to generate challenging prompts for the hard level, we experiment with various prompts to encourage GPT-3.5 to generate counterproductive sentences, as shown in Figure 3.

–**Compositionality.** We study three composition types, i.e., spatial, attribute-binding, and actions. The spatial prompts are collected using a pre-defined template, where a wide range of relations is utilized, e.g., "on the right of", "above", and "between." For attribute-binding, two attributes are exploited, i.e., colors and size. For each hardness level, the number of objects' compositions increased, ranging from 2 to 4. For the action-level compositional generation, we design prompts with multiple combinations of actions starting from ComCLIP [31]. We combine two sentences from ComCLIP [31] for the easy level. Then, for medium and hard levels, we randomly choose one sentence and feed 'Extend `text` to let the subject have at least three actions.' and 'Extend `text` with other subjects doing other actions' into GPT-3.5, respectively, to obtain the final prompt. Detailed examples are demonstrated in Figure 3.

–**Bias.** Random objects are sampled from LVIS datasets [26], combined with a pre-defined template, creating a meta-prompt. Then, the meta-prompt is fed to GPT-3.5 to produce the final prompt, as depicted in Figure 3. To ensure the prompts are agnostic towards the protected attributes, i.e., gender, race, and age, we manually validate them.

–**Fairness.** We adapt the bias prompts. For gender fairness, we replace gender-agnostic words, such as a person, with gender-specific words, such as man and woman. Whereas for style fairness, a style indicator is appended to the beginning of the bias prompt, as shown in Figure 3.

4. Evaluation for our Benchmark

As shown in Figure 4, we categorize the skills based on the evaluation criteria, one-to-many mapping. Thus the same skill may be assigned to several metrics.

4.1. Detection-Based Metrics

We utilize UniDet [76] for counting, spatial and attribute compositions because it supports a wide range of object classes, i.e., exceeding 700.

–**Counting.** We adopt the traditional detection measures, Precision, Recall, and F1-score, where Precision assesses the accuracy of additional objects, and Recall assesses the accuracy of missing objects.

–**Spatial Compositions.** Using a simple geometry module, we use the predicted bounding boxes to validate whether the spatial relation is grounded correctly. For instance, given the prompt "A cat above a car.", the predicted bounding boxes will be $\{x^1_{min}, y^1_{min}, x^1_{max}, y^1_{max}\}$

*https://platform.openai.com/docs/models/gpt-3-5

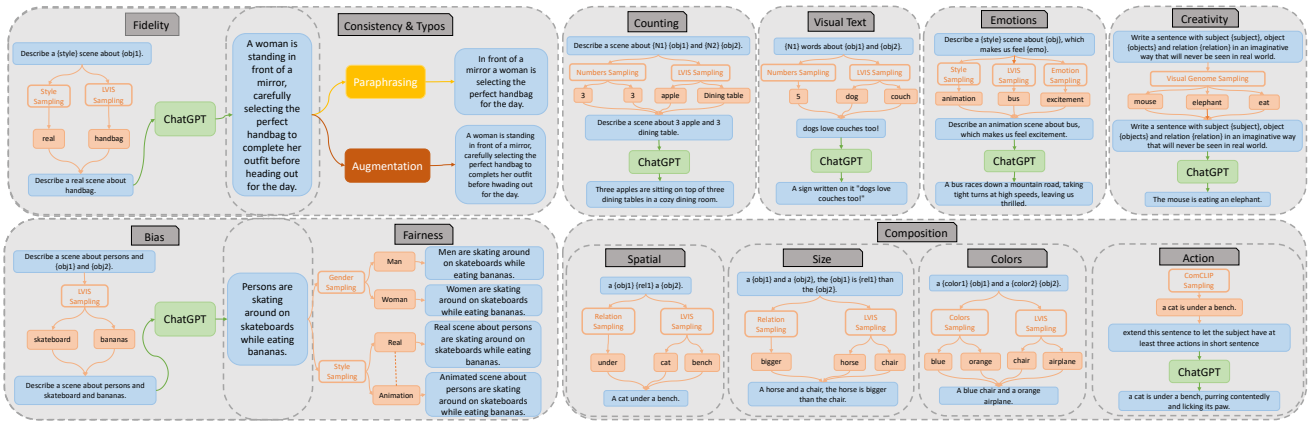

Figure 3: Our prompt generation pipeline. First, we create a meta-prompt, which is a template-based prompt (in blue). Then, we sample the skill-related attributes (in orange). Finally, we generate the final prompt using ChatGPT (in green).

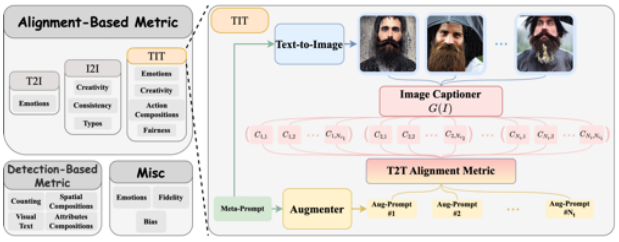

Figure 4: On the left is our evaluation taxonomy. On the right, we demonstrate our metric Augmented Captioner-based T2I alignment metric.

and $\{x^2_{min}, y^2_{min}, x^2_{max}, y^2_{max}\}$ for the cat and the car, respectively, and the grounded spatial relation is `above`. Then, our geometry module will assess whether the spatial relation, i.e., `above`, is grounded correctly based on the following condition: $(y^1_{min} < y^2_{min})$ or $(y^1_{max} < y^2_{max})$.

–**Attributes Compositions.** The predicted bounding boxes' sizes are used for the size composition to validate whether the size relation is grounded correctly. Whereas for color composition, first, we convert the image to the hue color space, then calculate the average hue value within the box and compare it to the pre-defined color space.

–**Visual Text.** We adopt Textsnake [39] and SAR [34] for text detection and recognition, respectively. The recognition accuracy is measured by the Character Error Rate (CER) [42] and the Normalized Edit Distance (NED) [65].

4.2. Alignment-Based Metrics

Three alignment paradigms are explored; Text-to-Image (T2I) (Sec. 4.2.1), Text-to-Image-to-Text (TIT) (Sec. 4.2.3), and Image-to-Image I2I (Sec. 4.2.4). In addition, we introduce our novel **A**ugmented **C**aptioner-based **T2I** Alignment metric, termed **AC-T2I** (Sec. 4.2.2).

4.2.1 T2I Alignment

One possible solution to assess the T2I model's grounding ability is measuring the text and image correlation, e.g., CLIPScore [27, 53]. While CLIP is widely used, its effectiveness has been repeatedly questioned, as it is not sensitive to fine-grained text-image alignment and fails to understand compositions [66, 73]. For instance, [73] shows that CLIP [53] can not distinguish between "the horse is eating the grass" and "the grass is eating the horse". This motivates us to propose our novel augmented captioner-based T2I alignment metric, termed AC-T2I, depicted in Figure 4.

4.2.2 AC-T2I Alignment Metric.

We propose a new T2I alignment metric, called AC-T2I, which overcomes the compositional relationship's limitations of existing large Vision-Language Models (VLMs) [31, 73], by utilizing the n-grams based metric, e.g., CIDEr [68] and BLEU [48]. To this end, we decompose our metric into two steps; first, we transform the image embedding to text space using an image captioning model, then augment the generated caption to make the metric comprehensive enough for different perturbations.

–**Reformatting T2I as TIT.** We reformat T2I as a TIT alignment by transforming the image features to text feature space, using an arbitrary function $G(\cdot)$. The function $G(\mathcal{I})$ could be interpreted as an image captioner, e.g., BLIP2 [35]. As shown in Figure 4, given a text prompt P^{org}, N_i images $\mathcal{I} = \{I_k\}_{k=1}^{N_i}$ are generated, which are fed to an image captioner $G(\mathcal{I})$ producing N_{c_i} captions $\mathcal{C} = \{C_k\}_{k=1}^{N_{c_i}}$, where N_{c_i} is the number of generated captions per image. Finally, the N_{c_i} captions are automatically evaluated using CIDEr [68] and BLEU [48] against the input prompt P^{org}.

–**Comprehensive TIT.** Instead of considering only the prompt P^{org} as the GT caption, N_t augmented prompts

$\mathcal{P}^{aug} = \{P_k^{aug}\}_{k=1}^{N_t}$ are generated using GPT-3.5, to measure the similarities comprehensively. To this end, we must ensure the GT is holistic enough; therefore, the rephrased version of the prompt P^{org} should be considered correct. Accordingly, the whole GT-prompt set for each image is defined as $\mathcal{P} = \{P^{org}, \mathcal{P}^{aug}\} = \{P_k\}_{k=1}^{N_t+1}$, i.e., one original prompt plus N_t augmented prompts.

Finally, for each prompt P^{org} we calculate the alignment score for each generated prompt-caption pair and select the highest score as the final alignment score, Eq. 1.

$$O_t = \frac{1}{N_i}\sum_{i=1}^{N_i} \max_{1\leq j\leq N_{c_i}, 1\leq k\leq N_t+1} S_t(C_{i,j}, P_k)), \quad (1)$$

where $S_t(,)$ is the text similarity scoring function, e.g., CIDEr [68], BLEU [48].

4.2.3 TIT Alignment

–**Emotions.** We explore a visual emotion classifier as illustrated in Sec. 4.3. Moreover, we apply our proposed metric, AC-T2I (Eq. 1), to avoid the aforementioned CLIP limitations, detailed in Section 4.2.1 and Section 4.2.2. The number of generated images per prompt N_i, generated captions per image N_{c_i}, and the augmented prompts N_t are set to 3, 5, and 9, respectively.

–**Creativity.** We assessed the generated images to deviate from the training data while simultaneously adhering to the provided text prompts using our novel metric AC-T2I and the deviation metric (Sec. 4.2.4). We set N_i and N_{c_i} to 3 and 5, respectively. Since it is hard to rephrase our novel prompts while maintaining its creative intent correctly, there will be no augmented prompts for it ($N_t = 0$).

–**Gender and styles fairness and action compositions.** The fairness score is defined as the disparities in subgroups' performance [24, 52], Eq. 2

$$Fairness_{score} = \frac{1}{N_s C_2}\sum_{i=1}^{N_s}\sum_{j=i+1}^{N_s} \frac{100 \times |A(i) - A(j)|}{max(A(i), A(j))}, \quad (2)$$

where $\frac{100}{N_s C_2 \times max(A(i),A(j))}$ is a normalization factor, N_s is the number of sub-groups, e.g., two for the gender and A is the accuracy measure, e.g., AC-T2I or CLIP scores. The less is, the better for $Fairness_{score}$. Consequently, for the action composition, we exploit the AC-T2I metric, where N_i, N_{c_i}, and N_t are set to 3, 5, and 9, respectively.

4.2.4 I2I Alignment

–**Creativity.** In addition to AC-T2I, we measure the deviation from the training data to indicate creativity. Accessing large models' training data is challenging, however, most of them are trained on LAION [63]. Accordingly, we use LAION image-text retrieval tools [8] to fetch training data, which search among the dataset using CLIP [53] and a KNN index to seek top-100 nearest images, denoted as $\mathcal{I}^{train}$ for each prompt. The deviation score is calculated based on Eq. 3.

$$\triangle(\mathcal{I}^{train}, I_i) = \frac{1}{2} - \frac{1}{2N_i}\sum_{i=1}^{N_i} S_v(\mathcal{I}^{train}, I_i), \quad (3)$$

where $S_v(,)$ is the visual similarity scoring function, i.e., CLIP [53], N_i number of generated images per prompt. The similarity can be regarded as the Nearest Neighbour (NN) distance from the training dataset, similar to [30].

–**Consistency and typos.** Given a prompt P^{org}, augmented prompt $\mathcal{P}^{aug}$ are generated using Parrot [14] for consistency and NLU-augmenter [19] for typos. Simultaneously, N_i images $\mathcal{I}$ and N_i augmented images $\mathcal{I}^{aug}$ are generated for P^{org} and $\mathcal{P}^{aug}$, respectively. Then the cosine similarity is calculated based on Eq. 4.

$$O_v = \frac{1}{2N_i}\sum_{i=1}^{N_i}\sum_{j=1}^{N_i} S_v(I_i, \mathcal{I}_j^{aug}), \quad (4)$$

where $S_v(,)$ is visual similarity scoring function; CLIP [53].

4.3. Miscellaneous

–**Emotion.** To comprehensively measure a T2I model's ability to generate images with grounded emotional tones, three evaluation metrics are proposed; AC-T2I (Sec. 4.2.3), T2I alignment (Sec. 4.2.1) and visual emotion classification accuracy. Regarding the visual emotion classifier, we train a ResNet-101 classifier based on combined datasets; FI [71] and ArtEmis [5], to ensure the model can handle diverse domains and scenarios.

–**Fidelity.** We rely on human evaluation, using Amazon Mechanical Turk (AMT) [13]. The annotators are asked to rate each image from 1-5, where 1 is the worst and 5 is the best. For a fair comparison, all the models' output images are shown in the same grid.

–**Bias.** Three bias attributes are assessed, i.e., gender, race, and age. First, the human faces are detected using ArcFace [15], and RetinaFace [16], then the facial attributes are detected using Dex [60]. Finally, the bias score is defined as the distribution skew, i.e., mean absolute deviation (MAD) [50]; Eq. 5, where the balanced case is $\frac{1}{N_b}$.

$$MAD = \frac{1}{N_b}\sum_{i=1}^{N_b}\left|\hat{N}_b - \frac{1}{N_b}\right|, \quad (5)$$

where N_b is the number of protected attribute groups, e.g., 2 genders, and $\hat{N}_b$ is the Dex output normalized count.

5. Experimental Results

5.1. Evaluated Methods

We comprehensively evaluate the performance of nine recent large-scale T2I models introduced as follows. *Transformer-based*: minDALL-E [2] and DALL-E-Mini [1]

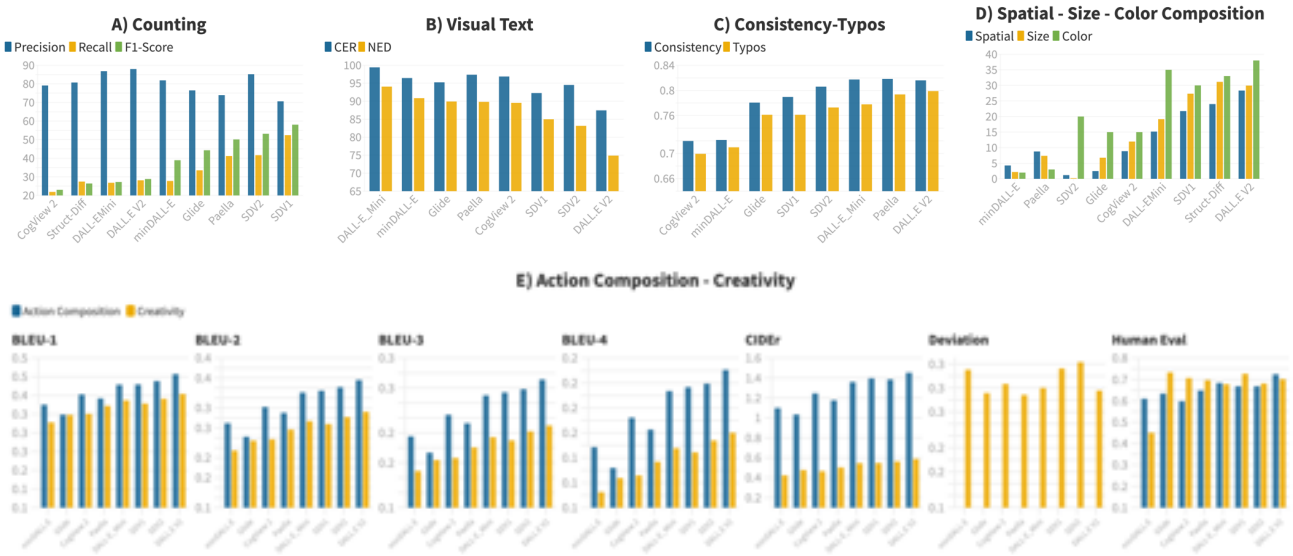

Figure 5: Quantitative results for nine skills are grouped into five sub-figures based on the evaluation criteria.

Figure 6: Qualitative results produced by DALLE-V2. Green and red boxes, respectively, frame the success and failure cases. More qualitative results for all models are demonstrated in the supplementary material.

are two different publicly available implementations of original DALL-E [56], which uses VQVAE [67] to encode images with grids of discrete tokens and a multimodal transformer for next token prediction. In addition, CogView2 [21] extend to a hierarchical transformer for fast super-resolution synthesis, and Paella [57] improve parallel token sampling based on MaskGIT [11]. *Diffusion-based*: GLIDE [45] and DALLE-V2 [55] decode images via diffusion with CLIP [53] embedding. Stable-Diffusion V1 [59] and V2 [3], (dubbed as SD-V1 and SD-V2) speed up the training of diffusion models [28] by leveraging the latent space of a powerful pre-trained VQVAE [67]. Finally, Struct-Diff [25] tackles the stable-diffusion compositions limitation by manipulating the cross-attention representations based on linguistic insights to preserve the compositional semantics.

5.2. Accuracy Skills Results

–**Counting.** We adopt the traditional detection measures, Precision, Recall, and F1-score. As shown in Figure 5 part A, DALLE-V2 [55] is the best in terms of precision. However, its recall is very poor, as it misses many objects. Whereas jointly considering recall and F1-score, SD-V1 [59] performs the best, despite its worst precision.

Finding #1. No agreement between precision and recall. We can select the appropriate model based on the application, which metric is preferred.

Finding #2. The more detailed prompt, the more accurate is counting performance. We explore three levels of prompts; 1) Vanilla prompt. The simplest form, e.g., two cups. 2) Meta-prompt. Intermediate level, e.g., describes a

scene containing two cups. 3) Detailed. The meta-prompt is fed to GPT-3.5 to generate a detailed description including the desired objects, e.g., two cups filled with hot coffee sitting side-by-side on a wooden table. In general, we may think that the simpler and straightforward prompts may lead to better results for the counting skill. Surprisingly, as shown in Figure 8, the Recall and F1-score always increase when the detailed prompt is used.

Finding #3. Composition-based solution is limited. We explore Struct-Diff [25] which tackle the compositionality limitation in SD-V1 [59]. As shown in Figure 5 part A, it increases the precision compared to SD-V1 [59]. However the recall and F1-score are decreased drastically.

–**Visual Text.** We utilize two text recognition metrics, CER [42] and NED [65], which are highly correlated (95%).

Finding #4. All models can not generate visual text even for the simplest case. As shown in Figure 5 part B, the best model is DALLE-V2 [55], which achieves a 75% error rate. However, the performances of all the models are far from acceptable, i.e., 10-20% error rate.

Finding #5. Confusion between picturing and writing. The models show a good language understanding of the mentioned semantics. However, they lean toward drawing them instead of visually writing them. For instance, in Figure 6, in the visual text column and first row, the model draws the "potted plant" instead of writing the words. Consequently, the model prefers to draw the "vessel" in the second row.

–**Emotion. Finding #6. All models suffer from generating emotion-grounded images.** Figure 10 shows the T2I and TIT alignment scores, i.e., BLEU [48], CIDEr [68], and CLIPScore [27], which are almost equally low and far from the acceptance range among the different models. To further validate our observation, we exploit an image-to-emotion classifier trained on combined datasets as discussed in Sec. 4.3. In addition, a human evaluation experiment is conducted. In both evaluations, the classifier and the human evaluation, we simplify the problem as a binary classification, where they are asked to classify the emotion, given the generated image, as a positive or negative emotion. Both report almost 50% accuracy across the entire models, precisely the random performance, where the number of classes is only two.

–**Fidelity.** We generate three distinct images using varying seeds for each of the models. Then, the annotators evaluate them on a scale of 1-5, where 5 is the best, and 1 is the worst. The normalized scores are reported in Figure 7. The best model is SD-V2 [3], where it achieves 62.4% while the worst one is minDALL-E [2] which achieves 52.2%. However all the models are far away from the accepted threshold, i.e., 80% which corresponds to 4 on our rating system (1-5).

5.3. Robustness and Generalization Results

–**Consistency and typos.** We measure the alignment between the images generated from the original prompt and paraphrased or perturbed prompts using CLIPScore [27], as discussed in Section 4.2.4.

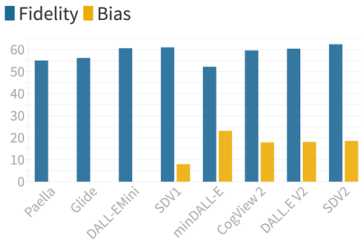

Figure 7: Quantitative results for Fidelity and Bias skills.

Finding #7. Models are robust against language perturbations. As shown in Figure 5 part C, all the models perform well against language perturbations and achieve between 70% to 82% alignment score. Specifically, DALLE-V2 [55] jointly achieves the best average performance for both skills, i.e., consistency and typos.

–**Spatial, Size, and Colors composition.**

Finding #8. The medium and hard levels are unattainable. For each skill, we define three hardness levels, i.e., easy, medium, and hard, and the reported results in the whole paper are the average accuracy for the three levels. However, we report only the easy level accuracy for the spatial, size, and color composition as the entire models suffer even from the easy level. Moreover, all models fails on the medium and hard levels, where they almost got zeros. Consequently, this raises the severe limitation of the models' composition ability. As shown in Figure 5 part D, the best model, DALLE-V2 [55], achieves 28.3%, 29.9%, and 38% for spatial, size, and colors composition, respectively.

Finding #9. Composition-based solution is limited. Similar to finding #3, we explore Struct-Diff [25], where it enhances the SD-V1 [59] performance by almost 3% on the easy level. However, it still fails on the challenging levels.

–**Action composition** Regarding the action compositionality, as illustrated in Figure 5 part E, DALLE-V2 [55] performs the best in generating compositions based on actions, according to both TIT alignment (i.e., highest CIDEr score 1.4538) and human evaluation result. Furthermore, all the scores align well with human evaluation results, confirming our metric's accuracy in evaluating this skill.

–**Creativity.** Since we retrieve the top-100 nearest training data with the CLIPScore[27], we obtain the average score of how the text prompt deviates from the training data, which is 0.4173. Since CLIP [53] maps images and text to shared space, the deviation score should be close if the generated image is aligned with the text. However, experimental results show that all the models fail to generate novel images, and the best model is SD-V2 [3], which achieves the highest deviation score of 0.3433. Due to creativity's very nature, it thrives on deviation. However, if the deviation becomes

Figure 8: Ablation study of the prompt details on the counting skill.

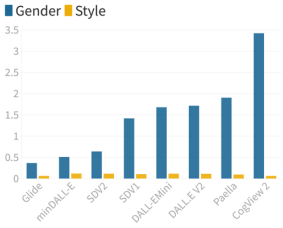

Figure 9: Gender fairness and style fairness results.

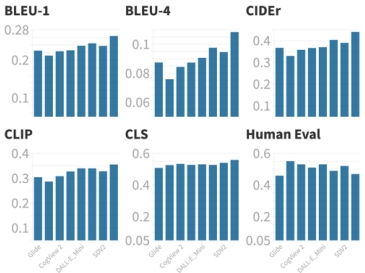

Figure 10: Emotions results using five different metrics, in addition to, the human evaluation.

excessive, the resulting generation may veer toward adverse hedonic and meaningless outcomes. Therefore, we also evaluate TIT (BLEU [48], CIDEr [68]) alignment to ensure the generation keeps the semantic meaning with the text prompt. For example, as Figure 5 part E) shows, Paella achieves a relatively high deviation score but performs poorly in terms of BLEU [48] and CIDEr [68], thus it deviates too much and lose original semantic meaning of corresponding text prompt, which is aligned with human evaluation result. Therefore, both metrics are indispensable for creativity evaluation, and more than one alone is required.

5.4. Fairness and Bias Results

Finding #10. The models are fair. As demonstrated in Figure 9, the maximum fairness score is 3.5% by Cogview 2 [21], which indicates that the difference in performance between sub-groups is negligible.

Finding #11. The models are slightly biased. In contrast to fairness, the models tend to be biased towards gender, as the average mean absolute deviation, Eq. 5, for DALLE-V2 [55], Cogview2 [21], and minDALLE [2] is 20%, as shown in Figure 7. However, SD-V1 [59] achieves the best results where the deviation is less than 8%. GLIDE [45], by design, is trained not to generate humans. Consequently, DALLE-Mini [1], and Paella [57] perform poorly regarding face generation. Therefore, GLIDE [45], DALLE-Mini [1], and Paella [57] are excluded from the bias measure.

5.5. Human Evaluation

To prove the effectiveness of our benchmark, we conduct a human evaluation using Amazon Mechanical Turk (AMT) [13] over 10% of our data across the entire skills. The human evaluation criteria are divided into two main groups: 1) Modular-based. The core blocks in each metric are evaluated. 2) End-to-End based. Using score-based evaluation.

–**Modular-based.** The UniDet [76] is the core block for counting, visual-text, spatial, color, and size composition. First, we ask humans to visually inspect its performance and report the true positives, false positives, and false positives. Then, we measure Person-correlation between the human F1-score and our F1-score, which shows a high correlation between our calculations and human evaluation, i.e., 93%.

Similarly, we measure the detection and recognition accuracy of Textsnake [39] and SAR [34], respectively. Again, the correlation is in our favor, 98% and 96%, respectively. Moreover, regarding the emotion, the annotators are asked to binary classify the emotions, i.e., positive or negative, given only the images. The results are highly aligned with our measure, where both agree that the models generate natural images, indicating that the prompt's emotion indicator is ignored. Regarding consistency and typos, the core module is the augmenter [14, 19]. To further ensure that all the generated prompts have the same meaning as the original prompt, we conduct a human study where we ask the users to rate the prompt pair (original and augmented) on a scale of 1 (not similar at all) to 5 (precisely same meaning). The results show a great alignment, i.e., 94%.

–**End-to-End based.** To assess the creativity metric, we equally select 100 images per model for each hardness level and let annotators score from 1 to 5 for each generated image considering: (1) whether the generated image is creative; (2) whether the image is aligned with the given prompt. For action composition skills, annotators are requested to assign a score between 1 to 5 based on the accuracy of the generated subject and actions in response to text prompts.

6. Conclusion

We introduce a comprehensive and reliable benchmark, dubbed HRS-Bench, for evaluating text-to-image (T2I) models. Our benchmark measures 13 skills, categorized into five major categories, and covers 50 applications, providing a holistic evaluation of T2I models. Through our evaluation of nine recent large-scale T2I models, we have identified areas where state-of-the-art models struggle to tackle these skills, highlighting the need for continued research and development. Our human evaluation results confirm the effectiveness and reliability of our benchmark. Further, our benchmark will help ease future T2I research and progress on improving the skills covered in this benchmark.

References

[1] Dallemini. https://github.com/borisdayma/dalle-mini. 2, 6, 9

[2] mindalle. https://github.com/kakaobrain/minDALL-E. 2, 6, 8, 9

[3] Stable-diffusion-v2. https://github.com/Stability-AI/stablediffusion. 2, 7, 8

[4] Panos Achlioptas, Maks Ovsjanikov, Leonidas Guibas, and Sergey Tulyakov. Affection: Learning affective explanations for real-world visual data. *arXiv preprint arXiv:2210.01946*, 2022. 3

[5] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. 3, 6

[6] Motasem Alfarra, Juan C. Pérez, Anna Frühstück, Philip H. S. Torr, Peter Wonka, and Bernard Ghanem. On the robustness of quality measures for gans, 2022. 2

[7] Shane Barratt and Rishi Sharma. A note on the inception score, 2018. 2

[8] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022. 6

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[10] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1, 2

[11] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 7

[12] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. 1, 2, 3

[13] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Anol Bhattacherjee and Brian Fitzgerald, editors, *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 6, 9

[14] Prithiviraj Damodaran. Parrot: Paraphrase generation for nlu., 2021. 4, 6, 9

[15] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 6

[16] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 6

[17] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*, 2021. 3

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2

[19] Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. Nl-augmenter: A framework for task-sensitive natural language augmentation, 2021. 4, 6, 9

[20] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1, 2

[21] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2, 7, 9

[22] Tan M. Dinh, Rang Nguyen, and Binh-Son Hua. Tise: Bag of metrics for text-to-image synthesis evaluation, 2021. 2

[23] Tan M Dinh, Rang Nguyen, and Binh-Son Hua. Tise: Bag of metrics for text-to-image synthesis evaluation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 594–609. Springer, 2022. 1, 2, 3

[24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 3, 6

[25] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1, 2, 7, 8

[26] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4

[27] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5, 8

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 7

[29] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018. 1

[30] Divyansh Jha, Kai Yi, Ivan Skorokhodov, and Mohamed Elhoseiny. Creative walk adversarial networks: Novel art generation with probabilistic random walk deviation from style norms. 3, 6

[31] Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. Comclip: Training-free compositional image and text matching. *arXiv preprint arXiv:2211.13854*, 2022. 2, 4, 5

[32] Kiku Johnson. *Sexual orientation, gender identity, and expression affirming approach and expansive practices*. 2019. 3

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[34] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8610–8617, 2019. 5, 9

[35] 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 preprint arXiv:2301.12597*, 2023. 5

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[37] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 1

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[39] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018. 5, 9

[40] Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. *arXiv preprint arXiv:2211.10780*, 2022. 3

[41] Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume abs/2204.07660, 2022. 3

[42] Andrew Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. 01 2004. 5, 8

[43] Deana F Morrow and Lori Messinger. *Sexual orientation and gender expression in social work practice: Working with gay, lesbian, bisexual, and transgender people*. Columbia University Press, 2006. 3

[44] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4467–4477, 2017. 1

[45] 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. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2, 7, 9

[46] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1

[47] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language

models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 4

[48] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5, 6, 8, 9

[49] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 1, 2

[50] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. 6

[51] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022. 1, 2, 3

[52] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017. 3, 6

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 6, 7, 8

[54] Micah Rajunov and A Scott Duane. *Nonbinary: Memoirs of gender and identity*. Columbia University Press, 2019. 3

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 7, 8, 9

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 7

[57] Dominic Rampas, Pablo Pernias, Elea Zhong, and Marc Aubreville. Fast text-conditional discrete denoising on vector-quantized latent spaces. *arXiv preprint arXiv:2211.07292*, 2022. 1, 2, 7, 9

[58] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 1

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 7, 8, 9

[60] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 6

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1, 4

[63] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 6

[64] D.A. Silverstein and J.E. Farrell. The relationship between image fidelity and image quality. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 1, pages 881–884 vol.1, 1996. 3

[65] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 5, 8

[66] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2, 5

[67] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 7

[68] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5, 6, 8, 9

[69] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[70] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. 4

[71] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 6

[72] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[73] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language

models behave like bags-of-words, and what to do about it? *arXiv e-prints*, pages arXiv–2210, 2022. 2, 5

[74] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 1

[75] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *arXiv preprint arXiv:2302.03675*, 2023. 1, 2

[76] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7571–7580, 2022. 4, 9

[77] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 1, 2

[78] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 1