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

Eslam Mohamed Bakr¹, Pengzhan Sun² *, Xiaojian Shen ¹ *, Faizan Farooq Khan ¹ *, Li Erran Li³, Mohamed Elhoseiny¹
{eslam.abdelrahman, xiaojian.shen, faizan.khan, mohamed.elhoseiny}@kaust.edu.sa
pengzhan@comp.nus.edu.sg, lilimam@amazon.com
¹King Abdullah University of Science and Technology (KAUST) ²National University of Singapore ³AWS 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 Holistic, Reliable, and Scalable. 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].

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,
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 wide range of covered scenarios by our proposed benchmark, termed HRS-Bench, and their object distribution.

- DrawBench 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.

### 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Eval Type</th>
<th>Prompt Type</th>
<th>Hardness Levels</th>
<th># Annotators</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrawBench</td>
<td>Human</td>
<td>Auto</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>DALL-EVAL</td>
<td>Human</td>
<td>Auto</td>
<td>3730</td>
<td>6</td>
</tr>
<tr>
<td>HE-T2I</td>
<td>Human</td>
<td>Template</td>
<td>90</td>
<td>20</td>
</tr>
<tr>
<td>TISE [23]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>HRS-Bench (Ours)</td>
<td>13</td>
<td>45000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>
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.

-HTE-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 predefined 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 \( \{ x_{min}, y_{min}, x_{max}, y_{max} \} \)
we sample the skill-related attributes (in metric, termed reduce our novel and Image-to-Image I2I (Sec. 4.2. Alignment-Based Metrics

accuracy is measured by the Character Error Rate (CER) and compare it to the pre-defined color space. 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. 4.2.2), and Image-to-Image I2I (Sec. 4.2.4). In addition, we introduce our novel Augmented Captioner-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 $G(Z)$ 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 $I = \{I_k\}_{k=1}^{N_i}$ are generated, which are fed to an image captioner $G(Z)$ producing $N_c$ captions $C = \{C_k\}_{k=1}^{N_c}$, where $N_c$ is the number of generated captions per image. Finally, the $N_c$ 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, $N_i$ augmented prompts

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.
where \( \mathcal{P}^{aug} = \{P_k^{aug}\}_{k=1}^{N_i} \) 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 \( \mathcal{P}^{org} \) should be considered correct. Accordingly, the whole GT-prompt set for each image is defined as \( \mathcal{P} = \{\mathcal{P}^{org}, \mathcal{P}^{aug}\} = \{P_k\}_{k=1}^{N_i+1} \), i.e., one original prompt plus \( N_i \) augmented prompts.

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

\[
O_i = \frac{1}{N_i} \sum_{i=1}^{N_i} \max_{1 \leq j \leq N_i, 1 \leq k \leq N_i+1} S_t(C_{i,j}, P_k),
\]

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

### 4.2.3 T2I 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 \( N_c \), and the augmented prompts \( N_i \) 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 \( N_c \) 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 (\( N_i = 0 \)).

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

\[
Fairness_{score} = \frac{1}{N_c N_i} \sum_{i=1}^{N_i} \sum_{j=1}^{N_c} \frac{100 \times |A(i) - A(j)|}{\max(A(i), A(j))},
\]

where \( \frac{100}{N_c N_i \max(A(i), A(j))} \) is a normalization factor, \( N_c \) 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 \( N_i \), \( N_c \), and \( N_i \) 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.

\[
\Delta(\mathcal{I}^{train}, I_i) = \frac{1}{2} - \frac{1}{2 N_i} \sum_{i=1}^{N_i} S_v(\mathcal{I}^{train}, I_i),
\]

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 \( \mathcal{P}^{org} \), 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 \( \mathcal{P}^{org} \) and \( \mathcal{P}^{aug} \), respectively. Then the cosine similarity is calculated based on Eq. 4.

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

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 \( \frac{1}{N_b} \).

\[
MAD = \frac{1}{N_b} \sum_{i=1}^{N_b} \left| \hat{N}_b - \frac{1}{N_b} \right|
\]

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]
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] extends 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.

**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 fail 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 color 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 CLIPScorescore[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
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 DALL-E 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, DALL-E-Mini [1], and Paella [57] perform poorly regarding face generation. Therefore, GLIDE [45], DALL-E-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


[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
[73] Mert Yuksekogul, 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?


