Evaluating and Improving Compositional Text-to-Visual Generation

Baiqi Li[1*] Zhiqiu Lin[1,2*] Deepak Pathak[1] Jiayao Li[1] Yixin Fei[1] Kewen Wu[1]
Xide Xia[2†] Pengchuan Zhang[2†] Graham Neubig[1†] Deva Ramanan[1†]
[1]CMU [2]Meta

Abstract

While text-to-visual models now produce photo-realistic images and videos, they struggle with compositional text prompts involving attributes, relationships, and higher-order reasoning such as logic and comparison. In this work, we conduct an extensive human study on **GenAI-Bench** *to evaluate the performance of leading image and video generation models in various aspects of compositional text-to-visual generation. We also compare automated evaluation metrics against our collected human ratings and find that* **VQAScore** *– a metric measuring the likelihood that a VQA model views an image as accurately depicting the prompt – significantly outperforms previous metrics such as CLIPScore. In addition, VQAScore can improve generation in a black-box manner (without finetuning) via simply ranking a few (3 to 9) candidate images.* **Ranking by VQAScore** *is 2x to 3x more effective than other scoring methods like PickScore and ImageReward at improving human ratings for DALL-E 3 and Stable Diffusion, especially on compositional prompts that require advanced visio-linguistic reasoning. Lastly, we identify areas for improvement in VQAScore, such as addressing fine-grained visual details. Despite mild limitations, VQAScore serves as the best automated metric as well as reward function for improving prompt alignment. We will release over 80,000 human ratings to facilitate scientific benchmarking of both generative models and automated metrics.*

1. Introduction

State-of-the-art text-to-visual models like Stable Diffusion [56], DALL-E 3 [2], Gen2 [17], and Sora [63] generate images and videos with exceptional realism and quality. Due to their rapid advancement, traditional evaluation metrics and benchmarks (e.g., FID scores on COCO [23, 37] and CLIPScores on PartiPrompt [22, 80]) are becoming insufficient [40, 50]. For instance, benchmarks should include more real-world *compositional* text prompts [43] that involve attribute bindings, object relationships, and logical reasoning, among other visio-linguistic reasoning skills (Figure 1). Moreover, it's crucial for automated evaluation metrics to measure how well the generated images (or videos) *align* with such compositional text prompts. Yet, widely used metrics like CLIPScore [22] function as *bag-of-words* [39, 69, 81] and cannot produce reliable alignment (faithfulness [25]) scores. Therefore, to guide the scientific benchmarking of generative models, we conduct a comprehensive evaluation of compositional text-to-visual generation alongside automated alignment metrics [5, 22, 30].

Evaluating text-to-visual generation. We collect a new text-to-visual benchmark, **GenAI-Bench**, which consists of 1,600 challenging real-world text prompts sourced from professional designers. Compared to benchmarks [25, 30, 45] such as PartiPrompt [80] and T2I-CompBench [26] (see Table 1), GenAI-Bench captures a wider range of aspects in compositional text-to-visual generation, ranging from **basic** (scene, attribute, relation) to **advanced** (counting, comparison, differentiation, logic). We collect a total of 38,400 human alignment ratings (1-to-5 Likert scales [50]) on images and videos generated by ten leading models, such as Stable Diffusion [56], DALL-E 3 [2], Midjourney v6 [48], Pika v1 [52], and Gen2 [17]. Our human study shows that while these models can often accurately generate basic compositions (e.g., attributes and relations), they still struggle with advanced reasoning (e.g., logic and comparison) (Figure 2). For instance, for "basic" prompts that do not require advanced reasoning, the state-of-the-art DALL-E 3 (most preferred by humans) achieves a remarkable average rating of 4.3, meaning its images range from having "*a few minor discrepancies*" to '*matching exactly*" with the prompts. However, its rating on "advanced" prompts drops to 3.4, indicating "*several discrepancies*". Figure 3 presents all human ratings.

Evaluating automated metrics. We also use the human ratings to benchmark automated metrics (e.g., CLIPScore [22], PickScore [30], and Davidsonian [5]) that measure the alignment between an image and a text prompt. Specifically, we show that a simple metric, **VQAScore**, which computes the likelihood of generating a "Yes" an-

*Co-first authors; †Co-senior authors.

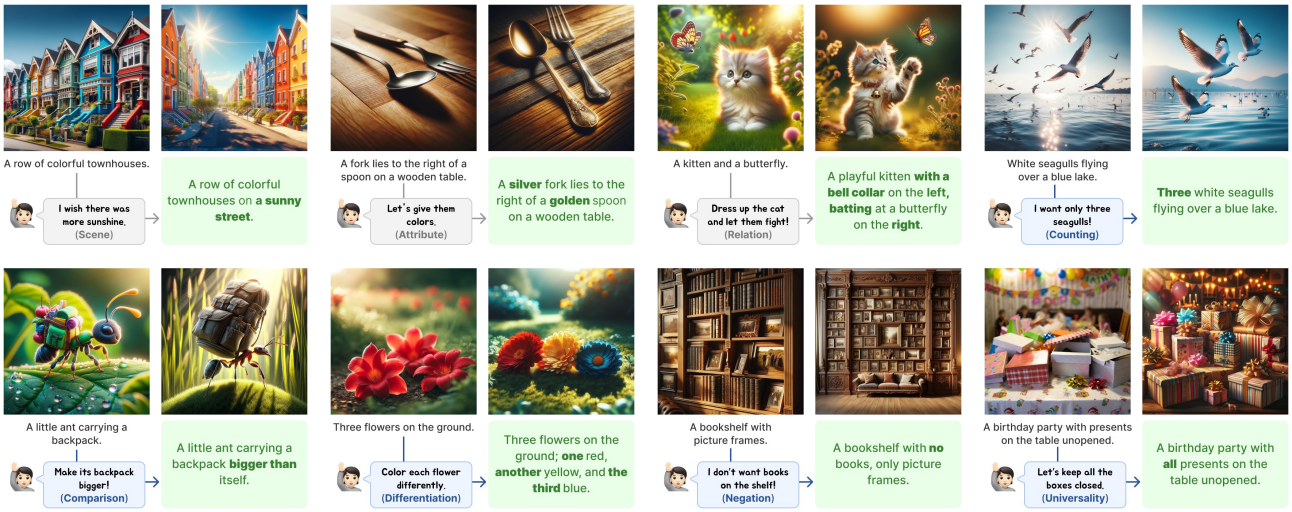

Figure 1. **Compositional text prompts of our GenAI-Bench (highlighted in green)** reflect how real-world users may seek precise control in text-to-visual generation. For example, users might add details by specifying compositions of basic visual entities and properties (highlighted in gray), such as scenes, attributes, and relationships (spatial/action/part). Moreover, user prompts may require advanced visio-linguistic reasoning (highlighted in blue), such as counting, comparison, differentiation, and logic (negation/universality). Appendix B details these essential skills with additional examples. Table 1 compares GenAI-Bench with previous benchmarks [26, 30, 58, 80].

swer to a question like "Does this figure show {text}?" from a VQA model, significantly surpasses previous metrics in correlating with human judgments. VQAScore can be calculated end-to-end from off-the-shelf VQA models, without finetuning on human feedback [30, 78] or decomposing prompts into QA pairs [5, 79]. VQAScore is strong because it leverages the compositional reasoning capabilities of recent multimodal large language models (LLMs) [9, 41] trained for VQA. For instance, our study adopts the leading **CLIP-FlanT5** model [40], which follows the best training practices in the literature [8], e.g., using a bidirectional encoder that allows the image and question embeddings to attend to each other. VQAScore based on CLIP-FlanT5 sets a new state-of-the-art on both GenAI-Bench and previous benchmarks like TIFA160 [25] and Winoground [69]. As such, we recommend adopting VQAScore over the "bag-of-words" CLIPScore, which has been overly abused in our community [28, 81]. We will release all human ratings to facilitate the development of automated metrics.

Improving generation with VQAScore. We show that one can improve text-to-image generation by ranking generated candidates with VQAScore and selecting the highest-scoring one. This ranking-based approach does not require any finetuning and can operate in a fully black-box manner [44], needing only an image generation API. Remarkably, simply ranking between 3 to 9 images can already enhance the average human ratings for DALL-E 3 and SD-XL by 0.2 to 0.3 (on a 1-to-5 Likert scale), setting the new closed-source and open-source SOTAs on GenAI-Bench. VQAScore significantly outperforms other metrics; for instance, using CLIPScore for ranking often leads to the same or lower human ratings. We present qualitative examples in Figure 7. Overall, VQAScore emerges as the most effective ranking metric, surpassing other metrics that rely on costly human feedback (e.g., PickScore [30]) or ChatGPT for prompt decomposition (e.g., Davidsonian [5]) by 2x to 3x.

Limitations. Lastly, we explore the implications of Goodhart's Law [18], particularly limitations of VQAScore in detecting fine-grained visual details and resolving linguistic ambiguity. Despite these mild limitations, we strongly urge the research community to adopt VQAScore as a reproducible supplement to non-reproducible human studies [50], or as a superior alternative to CLIPScore, which has ceased to be effective [28, 39, 81].

Contribution summary.

1. We conduct an extensive human study on text-to-visual generation using **GenAI-Bench**, revealing limitations of leading open-source and closed-source models.
2. By collecting over 80,000 human alignment ratings, we demonstrate that **VQAScore** is a simpler and more effective alternative to CLIPScore, which has been abused in current evaluations. We will release all human ratings to foster further research in this area.
3. We present a simple approach that improves generation by **ranking** images with VQAScore, significantly surpassing other scoring methods by 2x to 3x.

Table 1. **Comparing GenAI-Bench to existing text-to-visual benchmarks.** GenAI-Bench comprehensively covers essential aspects of compositional text-to-visual generation, emphasizing advanced reasoning skills (highlight in blue) that are required to parse real-world prompts. Moreover, GenAI-Bench tags each prompt with all evaluated aspects, in contrast to most benchmarks that assign merely one or two tags per prompt, even when multiple aspects are involved. GenAI-Bench also provides human ratings for both image and video generative models to support the benchmarking of automated metrics.

Benchmarks	Aspects Covered in Compositional Text-to-Visual Generation								Tagging	Human Annotation
	Scene	Attribute	Relation	Count	Negation	Universal	Compare	Differ		
PartiPrompt (P2) [80]	✓	✓	✓	✓	✓	✗	✗	✗	2 Tags	✗
DrawBench [58]	✓	✓	✓	✓	✗	✗	✗	✗	1 Tag	✗
EditBench [72]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
TIFAv1 [25]	✓	✓	✓	✓	✗	✗	✗	✗	All Tags	Images
Pick-a-pic [30]	✓	✓	✓	✓	✗	✗	✗	✗	✗	Images
T2I-CompBench [26]	✓	✓	✓	✓	✗	✗	✗	✗	1 Tag	Not Released
HPDv2 [77]	✓	✓	✓	✗	✗	✗	✗	✗	✗	Images
EvalCrafter [45]	✓	✓	✓	✓	✗	✗	✗	✗	✗	Videos
GenAI-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	All Tags	Images & Videos

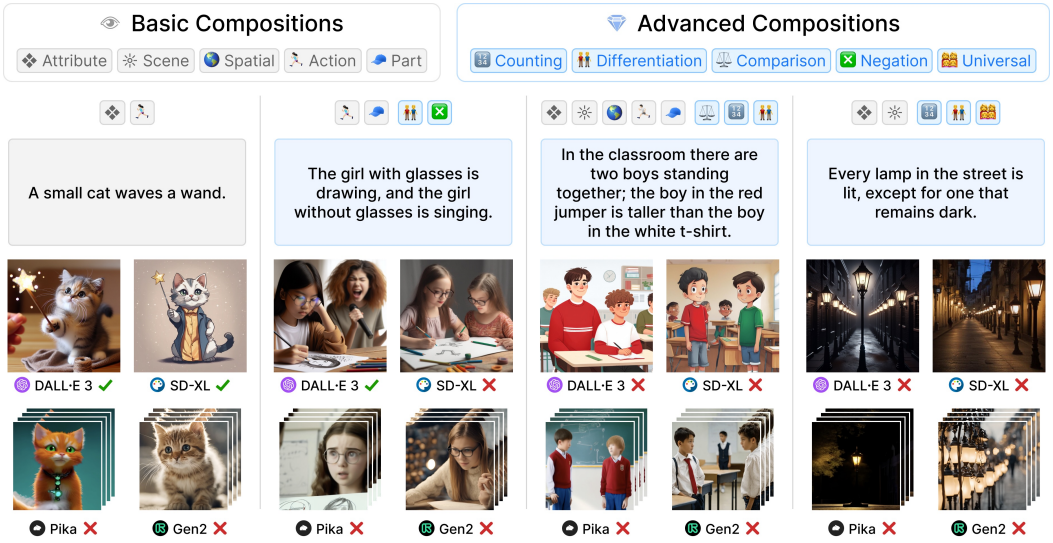

Figure 2. **GenAI-Bench challenges leading image and video generation models.** We present example prompts that fail even the state-of-the-art models such as DALL-E 3 [2], Stable Diffusion (SD-XL) [56], Pika [52], and Gen2 [17]. Note that each GenAI-Bench prompt is tagged with all evaluated aspects, allowing researchers to identify specific areas for improvement. In particular, "advanced" prompts (highlighted in blue) that require higher-order visio-linguistic reasoning – such as counting, comparison, differentiation, and logic – pose greater challenges to all generative models.

2. Related Works

Text-to-visual benchmarks. Early benchmarks mostly rely on captions from existing datasets like COCO [6, 25, 37, 55], focusing on generating simple objects, attributes, and scenes. Other benchmarks, such as HPDv2 [77] and Pick-a-pic [30], primarily evaluate image quality (aesthetic) using simpler text prompts. Recently, benchmarks like DrawBench [58], PartiPrompt [80], and T2I-CompBench [26] have shifted the focus to compositional text-to-image generation with an emphasis on attribute bindings and object relationships. Our GenAI-Bench escalates the challenge by incorporating real-world user prompts that require "advanced" reasoning (e.g., logic and comparison) to benchmark next-generation text-to-visual models.

Automated metrics. Perceptual metrics like IS [59], FID [23] and LPIPS [82] use pre-trained networks to assess the quality of generated imagery using reference images. To evaluate vision-language alignment (also referred to as faithfulness or consistency [10, 25, 46]), recent studies [4, 15, 16, 29, 32, 47, 57, 61, 74] primarily report CLIPScore [22], which measures (cosine) similarity of the embedded image and text prompt. However, CLIP cannot reliably process compositional text prompts due to its "bag-of-words" encoding [28, 39, 81]. Recent methods like Im-

ageReward [78], PickScore [30], and HPSv2 [77] further leverage human feedback to improve models like CLIP by finetuning on large-scale human ratings. Another popular line of works [7, 25, 26, 62, 75] uses LLMs like ChatGPT to decompose texts into simpler components for analysis, e.g., via question generation and answering (QG/A) [5]. For example, Davidsonian Scene Graph (or DSG) [5] decomposes a text prompt into simpler QA pairs and outputs a score as the accuracy of answers generated by a VQA model. However, Lin et al. [39, 40] show that such methods still face challenges in decomposing complex text prompts.

3. GenAI-Bench for Text-to-Visual Evaluation

In this section, we present **GenAI-Bench**, a challenging benchmark featuring real-world text prompts tagged with essential aspects of compositional text-to-visual generation.

Skill taxonomy. Prior literature on text-to-visual generation [26, 58, 80] focuses on generating "basic" objects, attributes, relations, and scenes. However, as illustrated in Figure 1, real-world prompts often require "advanced" compositional reasoning, including comparison, differentiation, counting, and logic. These "advanced" compositions extend beyond the "basic" ones. For example, real-world prompts may involve counting not just objects, but also attribute-object pairs and even object-relation-object triplets, e.g., "`three white seagulls flying over a blue lake`". Accordingly, we categorize compositional reasoning into "basic" (objects, scenes, attributes, and spatial/action/part relations) and "advanced" aspects (counting, comparison, differentiation, negation, and universality). Table 1 shows that GenAI-Bench uniquely covers all these essential aspects. We provide definitions and more examples in Appendix B.

GenAI-Bench. We collect 1,600 prompts from designers who routinely use text-to-image tools [48]. To improve diversity and quality, these designers also use ChatGPT for brainstorming prompt variants and correcting grammatical errors. Importantly, involving professional designers helps ensure the prompts are free from subjective or toxic content. For example, we observe that ChatGPT-generated prompts from T2I-CompBench [26] can include subjective (e.g., non-visual) phrases like "`a natural symbol of rebirth and renewal`". Similarly, Pick-a-pic [30] may contain inappropriate content (e.g., NSFW) crafted by malicious web users. We detail our collection procedure and discuss how we avoid these issues in the Appendix C. Lastly, we tag each prompt with *all* its evaluated aspects of compositional reasoning, in contrast to previous benchmarks that either release no tags [30, 45, 77] or limit them to one or two [26, 58, 80]. In total, GenAI-Bench provides over 5,000 human-verified tags with a roughly balanced distribution of skills. Specifically, about half of the prompts involve only "basic" compositions, while the other half poses greater challenges by incorporating both "basic" and "advanced" compositions. Figure 2 shows random prompts from GenAI-Bench that challenge some of the best generative models like DALL-E 3 [2] and Gen2 [17].

4. Human Evaluation via GenAI-Bench

We now present an extended human study of ten leading image and video generative models using GenAI-Bench.

Human evaluation. We evaluate six text-to-image models: Stable Diffusion [56] (SD v2.1, SD-XL, SD-XL Turbo), DeepFloyd-IF [12], Midjourney v6 [48], DALL-E 3 [2]; along with four text-to-video models: ModelScope [71], Floor33 [14], Pika v1 [52], Gen2 [17]. Next, we hire three annotators to collect 1-to-5 Likert scale human ratings for image-text or video-text alignment using the recommended annotation protocol of [50]:

> How well does the image (or video) match the description?
> 1. Does not match at all.
> 2. Has significant discrepancies.
> 3. Has several minor discrepancies.
> 4. Has a few minor discrepancies.
> 5. Matches exactly.

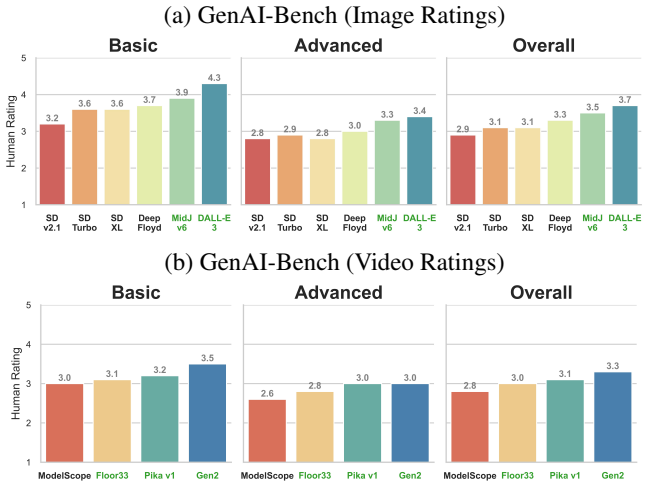

Figure 3. **Human evaluation on GenAI-Bench.** We show the average human alignment ratings on ten popular image and video generative models. We highlight closed-source models (e.g., DALL-E 3 [2]) in green. We find that (1) "advanced" prompts that require higher-order reasoning (e.g., negation and comparison) challenge all models more, (2) models using better text embeddings or captions (DeepFloyd-IF [12] and DALL-E 3 [2]) outperform others (SD-XL [56]), (3) open-source and video-generative models [17, 56] still lag behind their closed-sourced and image-generative counterparts, suggesting room for improvement.

Our collected human ratings indicate a high level of inter-rater agreement, with Krippendorff's Alpha reaching 0.72 for image ratings and 0.70 for video ratings, suggesting substantial agreement [25]. Specifically, the use of the Likert

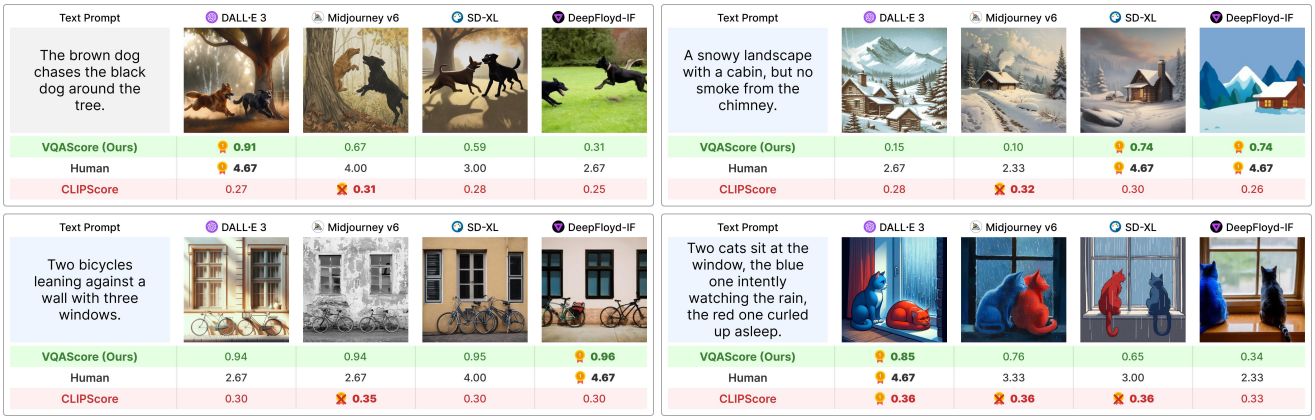

Figure 4. **VQAScore (based on CLIP-FlanT5 [40]) versus CLIPScore** on samples from GenAI-Bench. VQAScore shows a significantly stronger agreement with human ratings compared to CLIPScore [22], making it a more reliable tool for automatic text-to-visual evaluation, especially on real-world user prompts that involve complex compositional reasoning.

scale makes the final average rating interpretable. For example, a score near 5 implies that the model's generated images almost always "*match exactly*" with the input prompts.

Analysis. Figure 3 presents human ratings for basic, advanced, and overall prompts. Notably, advanced prompts that require complex visio-linguistic reasoning are much harder. For example, the top-performing DALL-E 3 scores 4.3 on basic prompts, indicating "*a few minor discrepancies*". However, its score drops to 3.4 on advanced prompts, indicating "*several minor discrepancies*". Interestingly, models (e.g., DeepFloyd-IF and DALL-E 3) using stronger text embeddings from LLMs (e.g., T5 [54]) outperform those using CLIP text embeddings (e.g., SD-XL). Lastly, we observe that open-source and video-generative models lag behind their closed-source and image-generative counterparts, suggesting avenues for future improvement. In Appendix C, we detail model performance across various aspects, highlighting challenges in higher-order reasoning like negation and comparison.

5. Evaluating Automated Metrics

We now use our human ratings to benchmark automated alignment metrics [5, 22, 79] on GenAI-Bench. We highlight a simple metric, **VQAScore**, as a superior alternative to the widely used CLIPScore.

VQAScore. Given an `image` and `text`, we calculate the probability of a "Yes" answer to a simple `question` like "Does this figure show '`{text}`'? Please answer yes or no.":

$$P(\text{"Yes"}|\texttt{image}, \texttt{question}) \quad (1)$$

We implement VQAScore using the state-of-the-art **CLIP-FlanT5** model [40] trained on 665K public VQA data [41]. For video-text pairs, we average the scores across all video frames following prior work [61]. We include more implementation details and pseudocode in Appendix D.

Table 2. **Evaluating the correlation of automated metrics with human ratings on GenAI-Bench.** We report Pairwise accuracy [13], Pearson, and Kendall, with higher scores indicating better performance for all. **VQAScore** based on the CLIP-FlanT5 VQA model [40] (detailed in Appendix E) achieves the strongest agreement with human ratings on images and videos, significantly surpassing popular metrics like CLIPScore [22], PickScore [30], and Davidsonian [5].

Method	GenAI-Bench (Image)			GenAI-Bench (Video)		
	Pairwise	Pearson	Kendall	Pairwise	Pearson	Kendall
CLIPScore [22]	51.9	19.3	13.5	53.6	25.3	18.0
BLIPv2Score [22]	55.1	25.0	20.7	54.6	25.3	20.1
ImageReward [78]	57.4	36.3	25.2	60.0	42.9	31.4
PickScore [30]	57.7	36.6	25.9	56.8	34.6	24.8
HPSv2 [77]	50.1	15.1	10.3	50.6	17.5	12.1
VQ2 [79]	52.5	16.2	14.8	52.8	18.0	15.5
Davidsonian [5]	54.2	32.5	23.1	55.9	32.3	23.5
VQAScore	**63.1**	**46.0**	**37.1**	**63.2**	**50.6**	**38.2**

Evaluation setup. To evaluate automated metrics on GenAI-Bench, we follow TIFA160 [25] to report the Pearson and Kendall coefficients, which reflect the correlation of the metric score with human judgment. However, Deutsch et al. [13] (EMNLP'23 outstanding paper) note several issues with these metrics. For example, Pearson assumes a linear relationship between metric and human scores, while Kendall ignores ties common in 1-to-5 Likert scales. As such, we also report **Pairwise** accuracy [13], designed to address these issues. We direct readers to [13] for equations and provide a brief overview below. For a dataset with M items (e.g., image-text pairs), there are two M-size score vectors: one for human ratings and one for metric scores. Pairwise accuracy (a value between 0 and 1) evaluates the percentage of agreement across all M x M pairs of items, i.e., if one item scores higher, lower, or ties with another

item in both human and metric scores. Additionally, we apply the tie calibration technique from [13] to find the optimal tie threshold for each metric.

Results. Table 2 shows that VQAScore significantly outperforms previous metrics such as CLIPScore [22], models trained with extensive human feedback [30, 77, 78], and QG/A methods that use the same CLIP-FlanT5 VQA model [5, 79]. Appendix G shows that VQAScore achieves the state-of-the-art performance on seven more alignment benchmarks such as TIFA160 [25] and Winoground [69]. Figure 4 qualitatively compare VQAScore against CLIPScore on random samples from GenAI-Bench. The strong performance of VQAScore makes it a more reliable tool for the future automated evaluation of text-to-visual models.

6. Improving Text-to-Visual Generation

VQAScore's superior performance in evaluating text-to-visual generation suggests its potential to improve generation as well. We now show that VQAScore can improve the alignment of DALL-E 3 [2] and SD-XL [56] by simply ranking candidate images.

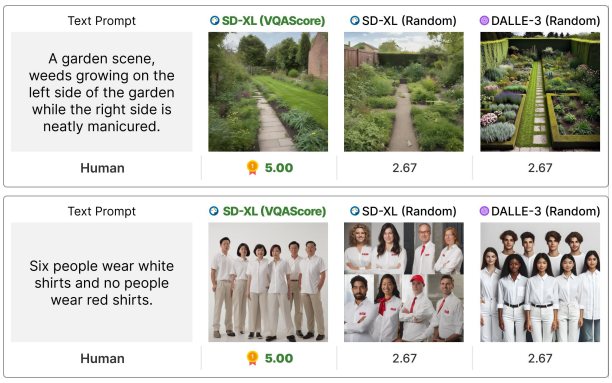

Figure 5. **VQAScore can select images generated by SD-XL [56] that outperform DALL-E 3's [2].** Although less powerful in prompt alignment than DALL-E 3, SD-XL [56] can still be improved by selecting three to nine candidate images with the highest VQAScore. We provide examples of how VQAScore ranks SD-XL images in Appendix A.

Ranking images by VQAScore. Given the same prompt, most text-to-visual models produce vastly different images in each inference run, with some being better than others. As such, we propose a *black-box* method [44] that improves text-to-image generation by ranking a few candidate images with VQAScore and selecting the highest-scoring one. This ranking-based approach is simple yet surprisingly effective. For instance, despite SD-XL's weaker prompt alignment compared to DALL-E 3, Figure 5 shows how VQAScore can select the best SD-XL images (from a few candidates) that outperform DALL-E 3's. Figure 7 shows that VQAScore can also improve the closed-source

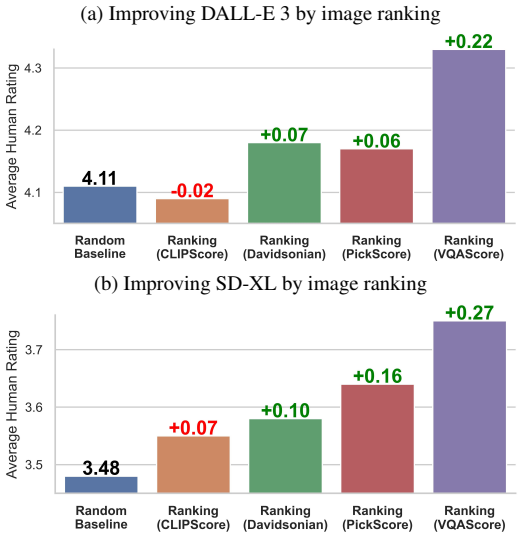

Figure 6. **Improving text-to-visual generation by ranking nine candidate images.** We show the performance gains over the *Random* baseline (no ranking) in green and decreases in red. Notably, selecting the highest-VQAScore images from nine candidates can significantly boost the overall human alignment ratings. In contrast, ranking by CLIPScore [22] results in the same or lower performance. Overall, VQAScore is 2x to 3x more effective than other methods that rely on costly human feedback (PickScore [30]) or decompose texts using ChatGPT (Davidsonian [5]). Table 3 details performance gains for more scoring methods across basic, advanced, and all prompts.

(black-box) DALL-E 3 by correctly selecting the most prompt-aligned images from three candidates.

A benchmark for ranking. To compare against other ranking metrics (e.g., CLIPScore and PickScore), we hire three annotators to rate nine generated images for each prompt. In this preliminary study, we randomly select 800 prompts from GenAI-Bench and collect 43,200 human ratings for 14,400 images generated by DALL-E 3 and SD-XL. We will release this benchmark for reproducibility and to facilitate the evaluation of future ranking metrics.

VQAScore achieves superior performance gains. Figure 6 confirms that ranking by VQAScore delivers the most significant improvements in human ratings. While ranking by CLIPScore [22] results in the same or even lower performance, VQAScore consistently improves with more images to rank. VQAScore is also 2x to 3x more effective than other ranking metrics that rely on expensive human feedback (e.g., PickScore [30]) or decompose texts via ChatGPT (e.g., Davidsonian [5]). Table 3 details the performance gains for ranking 3 to 9 images across basic, advanced, and all prompts. VQAScore notably improves the prompt alignment of DALL-E 3 and SD-XL by about 0.3 on "advanced" prompts that require complex visio-linguistic reasoning, such as counting, comparison, and logic.

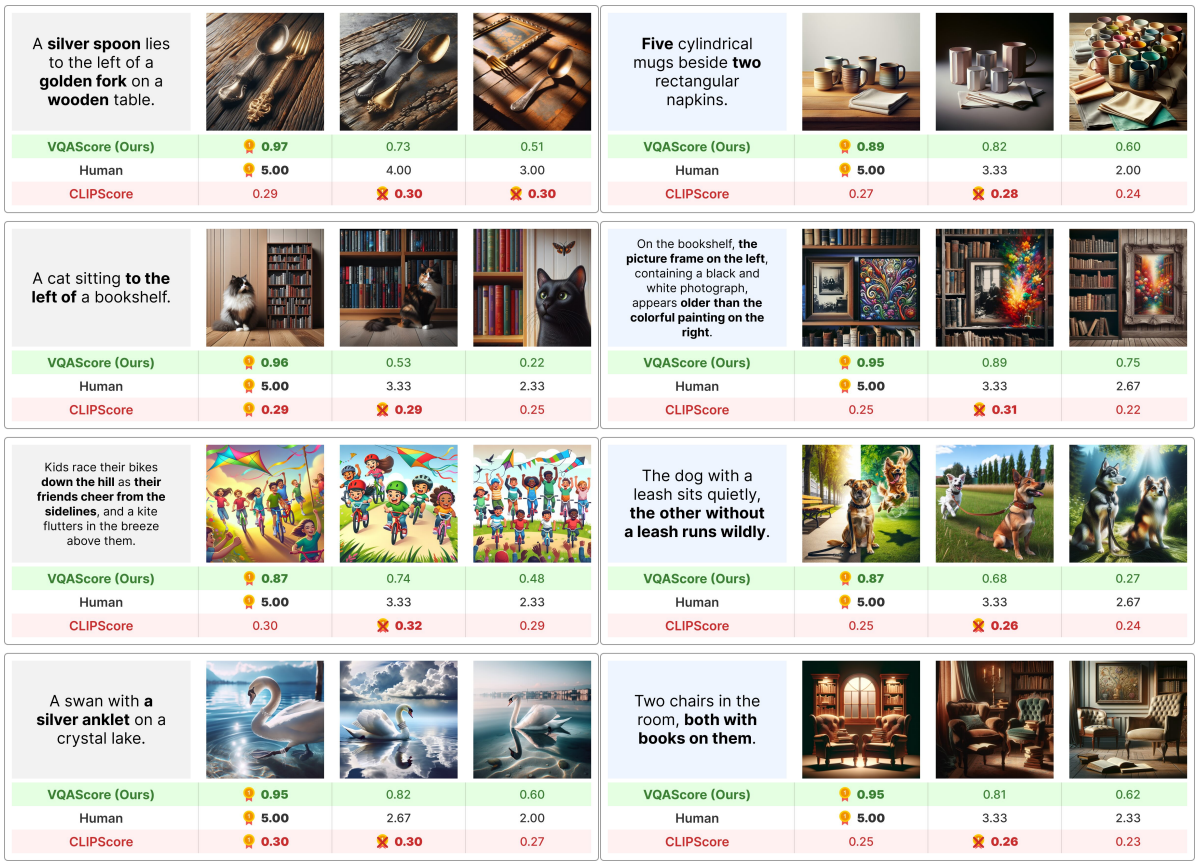

Figure 7. **Ranking DALL-E 3 generated images with VQAScore and CLIPScore.** VQAScore outperforms CLIPScore in ranking candidate images generated by DALL-E 3, particularly for prompts that involve attributes, relationships, and higher-order reasoning. This indicates that VQAScore can already improve text-to-image generation using only an image generation API [44]. We detail the performance gains achieved by VQAScore and other metrics in Figure 6 and Table 3.

Table 3. **Comparing scoring methods for image ranking.** We present the average human ratings of 7 popular scoring methods across basic, advanced, and all prompts on GenAI-Bench. Performance gains over the *Random* baseline (no ranking) are highlighted in green, while decreases are marked in red. Notably, some scoring methods like CLIPScore [22] can lead to a performance drop, particularly with an increasing number of images. For instance, CLIPScore results in a 0.04 drop when given more images to rank (from 3 to 9). In contrast, VQAScore demonstrates consistent and significant improvements with more images. VQAScore especially improves performance on the more challenging "advanced" prompts that require complex visio-linguistic reasoning skills like counting, comparison, and logic. For these "advanced" prompts, VQAScore boosts DALL-E 3 by 0.30 and SD-XL by 0.27 by ranking nine images, outperforming the second-best method PickScore [30] by 2x to 3x. For reference, we include human (oracle) performance (ranking by ground-truth human ratings). We will release over 40,000 human ratings to aid in the development of future ranking metrics.

Method	Basic		Advanced		Overall	
	3 Imgs	9 Imgs	3 Imgs	9 Imgs	3 Imgs	9 Imgs
Random	4.62	4.62	3.82	3.82	4.11	4.11
Human Oracle	$4.85_{+.23}$	$4.94_{+.32}$	$4.25_{+.43}$	$4.53_{+.71}$	$4.46_{+.35}$	$4.68_{+.57}$
CLIPScore [22]	$4.64_{+.02}$	$4.64_{+.02}$	$3.84_{+.02}$	$3.78_{-.04}$	$4.13_{+.02}$	$4.09_{-.02}$
ImageReward [78]	$4.66_{+.04}$	$4.60_{-.02}$	$3.88_{+.06}$	$3.90_{+.08}$	$4.16_{+.05}$	$4.15_{+.04}$
PickScore [30]	$\mathbf{4.68}_{+.06}$	$\mathbf{4.71}_{+.09}$	$3.87_{+.05}$	$3.87_{+.05}$	$4.16_{+.05}$	$4.17_{+.06}$
HPSv2 [77]	$4.66_{+.04}$	$4.68_{+.06}$	$3.86_{+.04}$	$3.83_{+.01}$	$4.14_{+.03}$	$4.13_{+.02}$
VQ2 [79]	$4.65_{+.03}$	$4.67_{+.05}$	$3.85_{+.03}$	$3.85_{+.03}$	$4.14_{+.03}$	$4.14_{+.03}$
Davidsonian [5]	$4.67_{+.05}$	$\mathbf{4.71}_{+.09}$	$3.88_{+.06}$	$3.89_{+.07}$	$4.16_{+.05}$	$4.18_{+.07}$
VQAScore	$\mathbf{4.68}_{+.06}$	$\mathbf{4.71}_{+.09}$	$\mathbf{3.98}_{+.16}$	$\mathbf{4.12}_{+.30}$	$\mathbf{4.23}_{+.12}$	$\mathbf{4.33}_{+.22}$

(a) Improving DALL-E 3 by image ranking

Method	Basic		Advanced		Overall	
	3 Imgs	9 Imgs	3 Imgs	9 Imgs	3 Imgs	9 Imgs
Random	4.05	4.05	3.17	3.17	3.48	3.48
Human (Oracle)	$4.41_{+.36}$	$4.62_{+.57}$	$3.53_{+.36}$	$3.84_{+.67}$	$3.84_{+.36}$	$4.12_{+.64}$
CLIPScore [22]	$4.11_{+.06}$	$4.17_{+.12}$	$3.21_{+.04}$	$3.21_{+.04}$	$3.53_{+.05}$	$3.55_{+.07}$
ImageReward [78]	$\mathbf{4.19}_{+.14}$	$4.21_{+.16}$	$3.25_{+.08}$	$3.29_{+.12}$	$3.59_{+.11}$	$3.62_{+.14}$
PickScore [30]	$4.18_{+.13}$	$4.24_{+.19}$	$3.28_{+.11}$	$3.31_{+.14}$	$3.60_{+.12}$	$3.64_{+.16}$
HPSv2 [77]	$4.15_{+.10}$	$4.22_{+.17}$	$3.25_{+.08}$	$3.29_{+.12}$	$3.57_{+.09}$	$3.62_{+.14}$
VQ2 [79]	$4.08_{+.03}$	$4.13_{+.08}$	$3.21_{+.04}$	$3.23_{+.06}$	$3.52_{+.04}$	$3.55_{+.07}$
Davidsonian [5]	$4.10_{+.05}$	$4.15_{+.10}$	$3.21_{+.04}$	$3.26_{+.09}$	$3.53_{+.05}$	$3.58_{+.10}$
VQAScore	$\mathbf{4.19}_{+.14}$	$\mathbf{4.32}_{+.27}$	$\mathbf{3.30}_{+.13}$	$\mathbf{3.44}_{+.27}$	$\mathbf{3.62}_{+.14}$	$\mathbf{3.75}_{+.27}$

(b) Improving SD-XL by image ranking

7. Goodhart's Law Still Applies

> *When a measure becomes a target, it ceases to be a good measure.*
>
> — Marilyn Strathern [64]

This quote conveys the essence of Goodhart's Law [18, 19]: an over-optimized metric inevitably loses its effectiveness. This phenomenon is well-documented in fields such as machine learning [24, 68], economics [11, 19], and education [3, 31]. Acknowledging that VQAScore is also subject to this law, we examine its limitations as an automated metric and suggest avenues for future improvements.

Limitations of VQAScore. We conduct a qualitative study by manually examining samples where VQAScore and human ratings disagree. Figure 8 identifies three failure cases: (1) miscounting when there are too many objects, (2) overlooking fine-grained visual details, and (3) misinterpreting linguistic ambiguity. We posit that VQA models with higher image resolution [60] and more capable language models [49, 67] may improve on these challenging aspects. Despite these mild limitations, we strongly recommend adopting VQAScore as a more reliable alternative to CLIPScore, which has already ceased to be an effective metric [28, 39, 81]. We believe VQAScore also serves well as a reproducible supplement to non-reproducible human studies [50].

8. Conclusion

Limitations and future work. Currently, GenAI-Bench does not evaluate several vital aspects of generative models [34, 45, 51, 76], such as toxicity, bias, aesthetics, and video motion. Although our ranking-based approach is effective, future work may explore white-box finetuning techniques for more efficient inference.

Summary. We have conducted an extensive human study with GenAI-Bench, focusing on both compositional text-to-visual generation and automated evaluation metrics. We show a straightforward ranking-based method that improves the prompt alignment of black-box generative models. By discussing Goodhart's Law, we hope to encourage further research into automated evaluation techniques, which is essential to the scientific progression of this field.

9. Acknowledgement

We express our deepest gratitude to the Meta GenAI team (Xiaoliang Dai, Miao Liu, Peizhao Zhang, Peter Vajda, Ning Zhang) for supporting this work. We thank Pengliang Ji, Zihan Wang, Jean de Dieu Nyandwi, Simran Khanuja, Zixian Ma, and Ranjay Krishna for their invaluable discussions during the development of this work. We also thank Tiffany Ling for her contributions to the visual design.

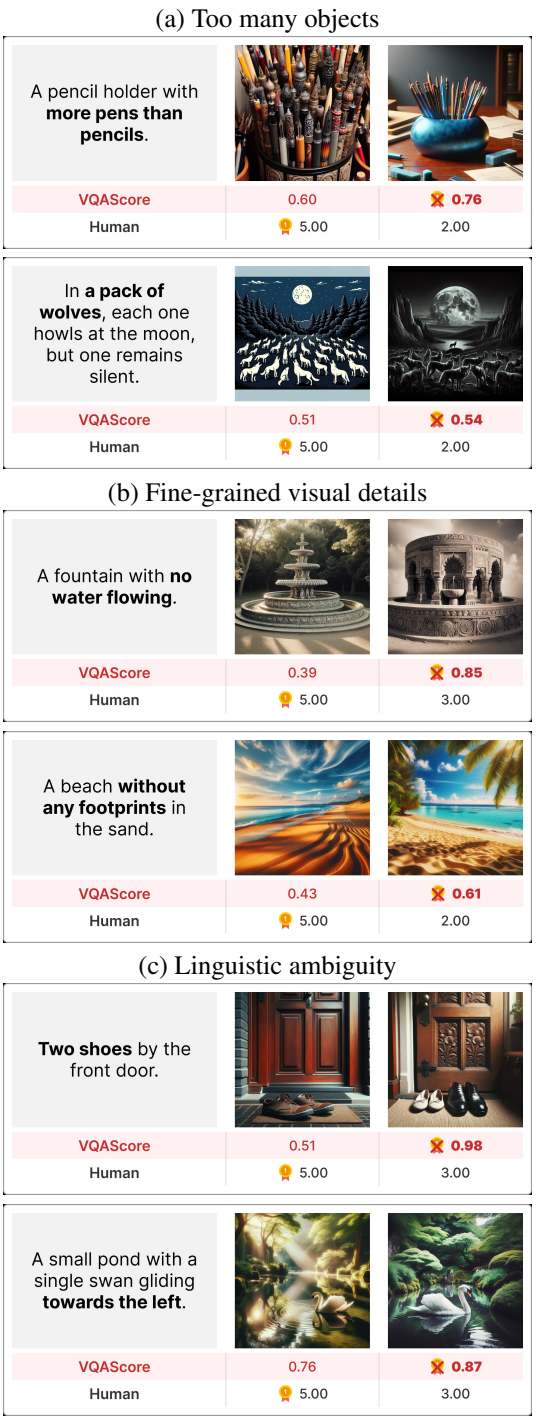

Figure 8. **Limitations of VQAScore (please zoom into the figures for a detailed view).** We identify three failure cases of VQAScore. (a) While VQAScore can reasonably count objects in small quantities, it struggles with larger numbers. (b) VQAScore can overlook small visual details, such as entities that occupy only a small portion of the image. (c) VQAScore may not understand ambiguous prompts, misinterpreting "two shoes" as "two pairs of shoes", or "towards the left (of the viewer)" as "towards the left (of the swan)".

References

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. 14, 16

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *https://cdn.openai. com/papers/dall-e-3.pdf*, 2023. 1, 3, 4, 6, 14, 18

[3] Mario Biagioli. Watch out for cheats in citation game. *Nature*, 535(7611):201–201, 2016. 8

[4] 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. 3

[5] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 1, 2, 4, 5, 6, 7, 14, 16, 18

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023. 3, 13, 18

[7] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023. 4

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 16

[10] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 3

[11] Jón Danıelsson. The emperor has no clothes: Limits to risk modelling. *Journal of Banking & Finance*, 26(7):1273–1296, 2002. 8

[12] Deepfloyd IF. Deepfloyd IF. https://github.com/deep-floyd/IF, 2024. 4, 14

[13] Daniel Deutsch, George Foster, and Markus Freitag. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, 2023. 5, 6, 18

[14] Floor33. Floor33. https://www.morphstudio.com/, 2023. 4, 14

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[16] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3

[17] Gen2. Gen2. https://research.runwayml.com/gen2, 2024. 1, 3, 4, 14

[18] Charles Goodhart. *Goodhart's law*. Edward Elgar Publishing Cheltenham, UK, 2015. 2, 8

[19] Charles AE Goodhart and CAE Goodhart. *Problems of monetary management: the UK experience*. Springer, 1984. 8

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 17

[21] Jianshu Guo, Wenhao Chai, Jie Deng, Hsiang-Wei Huang, Tian Ye, Yichen Xu, Jiawei Zhang, Jenq-Neng Hwang, and Gaoang Wang. Versat2i: Improving text-to-image models with versatile reward. *arXiv preprint arXiv:2403.18493*, 2024. 13

[22] 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. 1, 3, 5, 6, 7, 17, 18

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 3

[24] Jennifer Hsia, Danish Pruthi, Aarti Singh, and Zachary C Lipton. Goodhart's law applies to nlp's explanation benchmarks. *arXiv preprint arXiv:2308.14272*, 2023. 8

[25] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 1, 2, 3, 4, 5, 6, 13, 14, 16, 18

[26] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 1, 2, 3, 4, 13, 18

[27] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 17

[28] Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, 2023. 2, 3, 8, 17

[29] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[30] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. 1, 2, 3, 4, 5, 6, 7, 17, 18

[31] Vladlen Koltun and David Hafner. The h-index is no longer an effective correlate of scientific reputation. *PLoS One*, 16 (6):e0253397, 2021. 8

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[33] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 13

[34] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *arXiv preprint arXiv:2311.04287*, 2023. 8

[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. 17

[36] Jiachen Li, Weixi Feng, Wenhu Chen, and William Yang Wang. Reward guided latent consistency distillation. *arXiv preprint arXiv:2403.11027*, 2024. 13

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3, 18

[38] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models, 2023. 13

[39] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2024. 1, 2, 3, 4, 8, 17

[40] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 1, 2, 4, 5

[41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 5, 17

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 16

[43] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 1

[44] Shihong Liu, Zhiqiu Lin, Samuel Yu, Ryan Lee, Tiffany Ling, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*, 2024. 2, 6, 7

[45] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023. 1, 3, 4, 8

[46] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. 3

[47] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[48] Midjourney. Midjourney. https://www.midjourney.com, 2024. 1, 4, 13, 14

[49] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8

[50] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023. 1, 2, 4, 8

[51] Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. *arXiv preprint arXiv:2401.12425*, 2024. 8

[52] Pika. Pika. https://www.pika.art/, 2024. 1, 3, 4, 14

[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. 17

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 5

[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*, 1 (2):3, 2022. 3

[56] 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, 3, 4, 6, 14

[57] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3, 13

[58] 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. 2, 3, 4, 13, 18

[59] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[60] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*, 2024. 8

[61] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3, 5

[62] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *arXiv preprint arXiv:2307.04749*, 2023. 4

[63] Sora. Sora. https://openai.com/sora, 2024. 1

[64] Marilyn Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997. 8

[65] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. *arXiv preprint arXiv:2311.17946*, 2023. 18

[66] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. *arXiv preprint arXiv:2311.17946*, 2023. 13

[67] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022. 8

[68] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in neural information processing systems*, 33:407–417, 2020. 8

[69] 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. 1, 2, 6, 13, 17, 18

[70] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023. 13

[71] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 4, 14

[72] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 3

[73] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. *arXiv preprint arXiv:2303.14465*, 2023. 18

[74] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[75] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024. 4, 18

[76] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *arXiv preprint arXiv:2401.04092*, 2024. 8

[77] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 3, 4, 5, 6, 7, 17, 18

[78] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 2, 4, 5, 6, 7, 17, 18

[79] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023. 2, 5, 6, 7, 18

[80] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1, 2, 3, 4, 13, 18

[81] 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? In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 8, 17

[82] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3