Evaluating Text-to-Visual Generation with Image-to-Text Generation

Zhiqiu Lin[1,2], Deepak Pathak[1], Baiqi Li[1], Jiayao Li[1], Xide Xia[2], Graham Neubig[1], Pengchuan Zhang[2*], and Deva Ramanan[1*]

[1] Carnegie Mellon University
[2] Meta
*Co-Senior Authors
https://linzhiqiu.github.io/papers/vqascore

Abstract. Despite significant progress in generative AI, comprehensive evaluation remains challenging because of the lack of effective metrics and standardized benchmarks. For instance, the widely-used CLIPScore measures the alignment between a (generated) image and text prompt, but it fails to produce reliable scores for complex prompts involving compositions of objects, attributes, and relations. One reason is that text encoders of CLIP can notoriously act as a "bag of words", conflating prompts such as `"the horse is eating the grass"` with `"the grass is eating the horse"` [39, 65, 78]. To address this, we introduce the **VQAScore**, which uses a visual-question-answering (VQA) model to produce an alignment score by computing the probability of a `"Yes"` answer to a simple `"Does this figure show {text}?"` question. Though simpler than prior art, VQAScore computed with off-the-shelf models produces state-of-the-art results across many (8) image-text alignment benchmarks. We also compute VQAScore with an in-house model that follows best practices in the literature. For example, we use a bidirectional image-question encoder that allows image embeddings to depend on the question being asked (and vice versa). Our in-house model, **CLIP-FlanT5**, outperforms even the strongest baselines that make use of the proprietary GPT-4V. Interestingly, although we train with only images, VQAScore can also align text with video and 3D models. VQAScore allows researchers to benchmark text-to-visual generation using complex texts that capture the compositional structure of real-world prompts. Towards this end, we introduce **GenAI-Bench**, a more challenging benchmark with 1,600 compositional text prompts that require parsing scenes, objects, attributes, relationships, and high-order reasoning such as comparison and logic. GenAI-Bench also collects over 15,000 human ratings for leading image and video models such as Stable Diffusion, DALL-E 3, Midjourney, and Gen2. We open-source our data, model, and code at link.

Keywords: Vision-Language Models · Visio-Linguistic Compositionality · Evaluation of Generative Models

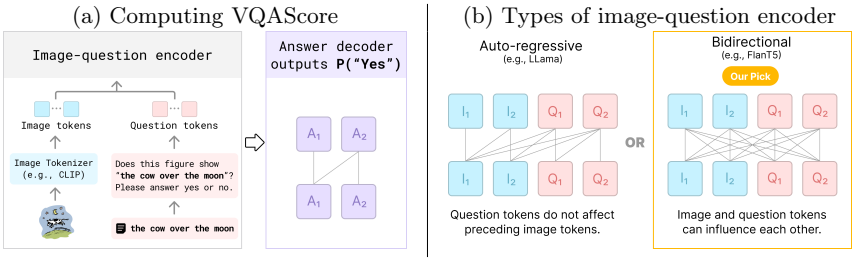

Fig. 1: VQAScore. Figure (a) computes the VQAScore between an image and text by first converting the text into the question "Does this figure show '{text}'? Please answer yes or no." The image and question (after tokenization) are then fed into an image-question encoder, followed by an answer decoder that outputs the probability of "Yes". Appendix C details the implementation and pseudocode. Our simple VQAScore based on off-the-shelf VQA models [9, 40] even rivals prior art that uses proprietary models [29, 70, 76] such as GPT4-Vision. **Figure (b)** highlights the architectural choice of the image-question encoder. While popular open-source VQA models such as LLaVA-1.5 [40] are derived from auto-regressive architectures like LLama-2 [66] where question tokens do not affect preceding image tokens, we find it beneficial to adopt bidirectional encoders, e.g., FlanT5 [8]. This allows the image to be "looked at" differently depending on the question, and vice versa. VQAScore based on our **CLIP-FlanT5** model achieves a new state-of-the-art across text-to-image/video/3D alignment benchmarks. Figure 2 shows examples of VQAScore's superior agreement with human judgments of images generated from complex text prompts.

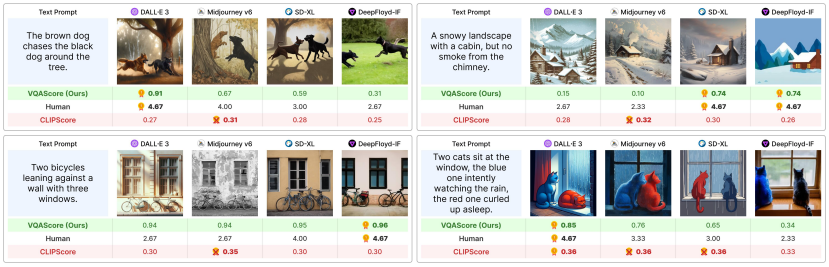

Fig. 2: VQAScore (based on CLIP-FlanT5) versus CLIPScore on samples from our GenAI-Bench (detailed in Section 5). GenAI-Bench consists of 1,600 text prompts spanning diverse compositional reasoning skills that challenge even leading models such as DALL-E 3 [1] and Stable Diffusion (SD) [56]. VQAScore shows a significantly stronger agreement with human judgments compared to CLIPScore [17], making it a more reliable tool for automatic text-to-visual evaluation. We open-source our code and models for VQAScore at link.

1 Introduction

Metrics play a key role in the evolution of science. For instance, perceptual metrics such as FID [18], IS [59], and LPIPS [80] have enabled tremendous

progress by allowing researchers to systematically assess the *quality* of generated imagery. However, the generative AI community still lacks a robust metric that reveals how well an image *aligns* with an input text prompt. Indeed, generative models such as DALL-E 3 [1] and Gen2 [15] produce remarkably photo-realistic images and videos that can still fail to align with input text prompts [21, 23, 28].

Challenges in evaluation. Contemporary generative models [1, 10, 58, 79] primarily rely on *subjective* human evaluation [42, 58, 60, 75, 77] which can be expensive and difficult to reproduce. For systematic benchmarking, recent work [2, 3, 36, 57, 60, 71] adopts metrics such as CLIPScore [17, 54], which measures the (cosine) similarity of the embedded image and text prompt. However, accurately measuring vision-language alignment remains a significant challenge for even leading vision-language models (VLMs), because it requires advanced *compositional* reasoning skills (that may be as difficult as the underlying generative task!). Studies [25, 39, 46, 68, 78] show that VLMs like CLIP struggle with compositional text prompts involving multiple objects, attribute bindings, spatial/action relations, counting, and logical reasoning. Given the current state of the art, the power of standard evaluation metrics lags far behind the power of the generative models that they are evaluating.

Decomposing texts via LLMs (prior art). Recent neuro-symbolic methods [6, 7, 16, 21, 63, 76] use off-the-shelf large language models (LLMs) like ChatGPT [49, 51] to tackle compositional reasoning through a *divide-and-conquer* approach, i.e., decomposing complex prompts into modular components. For example, visual programming [16, 63] uses LLMs to translate task instructions into symbolic programs, which themselves can invoke expert VLMs to return intermediate outputs like object counts [7]. This inspires many recent methods [7,23,70,76] to compute image-text alignment by decomposing the text prompt into simpler components, e.g., question-answer pairs. For example, TIFA [21] decomposes a prompt "`parent pointing at child`" into questions like "who is pointing at the child?" and "who is being pointed at?", and returns the accuracy score of the answers generated by a visual-question-answering (VQA) model. However, these approaches struggle with more compositional text prompts, e.g., those from challenging benchmarks such as Winoground [65]. For example, given a prompt "`someone talks on the phone happily while another person sits angrily`", the latest divide-and-conquer method Davidsonian [6] generates nonsensical questions like "is the someone happy?" and "is there another person?".

VQAScore (ours). Using recent VQA models based on multimodal LLMs [9, 41], we propose the following *end-to-end* approach that computes the generative likelihood [39] of an answer to a simple question (see Figure 1). Given an `image` and `text`, we define their alignment to be the following probability:

$$P(\text{"Yes"}|\texttt{image}, \text{"Does this figure show '}\texttt{\{text\}}\text{'? Please answer yes or no."}) \quad (1)$$

We term this approach **VQAScore**. Despite its simplicity, VQAScore implemented via open-source VQA models [9, 40] outperforms nearly all prior art including CLIPScore [17], models trained with extensive human feedback [28,73,74],

and divide-and-conquer methods [6, 7, 21, 76]. VQAScore even competes with approaches that rely on proprietary models [29, 70] like GPT4-Vision trained on much larger datasets. We evaluate across a comprehensive suite of alignment benchmarks including Winoground [65], EqBen [68], TIFA160 [21], Flickr8K [17], DrawBench [58], EditBench [67], COCO-T2I [38], and Pick-a-Pic [28]. We analyze the performance of various open-source models with respect to the benchmarks, and propose innovations in both modeling and benchmarking below.

What makes VQAScore effective? To isolate factors crucial for image-text alignment, we train in-house VQA models controlling for architectures, training data, and training recipes. Recall that VQA models need be trained on (image, question, answer) examples [40]. We first point out that image-text alignment requires models to expose answer likelihoods rather than simply generate answer tokens (as much past work does [6,21]). Another crucial architectural choice is the type of image-question encoder. Many popular VQA models (e.g., LLaVA [40,41]) are derived from next-token autoregressive LLMs (e.g., Llama-2 [66]) where question embeddings depend on previously-encoded image tokens, but *not* vice versa. These are often known as uni-directional "decoder-only" architectures. However, we find it beneficial to allow visual embeddings to be influenced by the question being asked (and vice versa). Indeed, there exists tremendous evidence from neuroscience that humans parse imagery differently depending on the prompted task (via top-down feedback [20]). We operationalize this via a bidirectional "encoder-decoder" language model, FlanT5 [8]. Specifically, we combine a pre-trained CLIP vision-encoder with a pre-trained FlanT5, which encodes image and question embeddings bidirectionally but generates answers auto-regressively (see Figure 1). By finetuning on public VQA datasets [40], our final **CLIP-FlanT5** sets a new state-of-the-art across all benchmarks. Interestingly, even though we need only simple question-answers at inference time (1), VQAScore likely benefits from FlanT5's strong reasoning ability, trained on more than 400 language datasets with challenging question-answer pairs [8].

GenAI-Bench. We find that popular benchmarks for generative models [23,28,58,70] like PartiPrompt [77] do not capture the compositional structure of real-world text prompts (e.g., Winoground [65]). To remedy this, we identify a set of crucial skills for text-to-visual generation, ranging from basic (object, scene, attribute, and relation understanding) to advanced (comparison, differentiation, logical reasoning, and counting). Figure 3 presents illustrative examples. Although these skills frequently appear in user prompts, we find that existing benchmarks [21, 23, 77] do not comprehensively cover them. To address the gaps, we introduce GenAI-Bench to evaluate both (1) text-to-visual generation models and (2) automated metrics. First, GenAI-Bench evaluates text-to-visual generation by collecting 1,600 prompts that cover essential visio-linguistic compositional reasoning skills. This allows us to identify the limitations of popular generative models such as Stable Diffusion, Midjourney, DALL-E 3, Pika, and Gen2. For quality purposes, the prompts are sourced from graphic designers who use text-to-visual tools in their profession. Next, GenAI-Bench evaluates automated metrics by collecting over 15,000 human ratings for ten leading text-to-visual

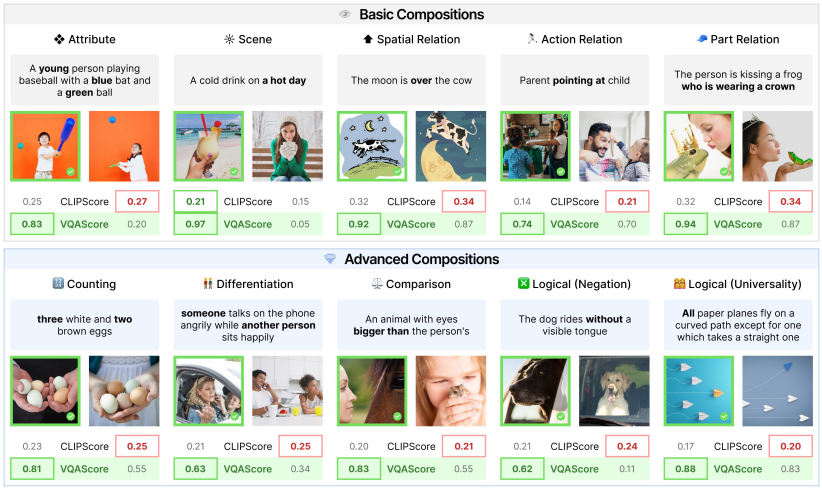

Fig. 3: VQAScore (based on CLIP-FlanT5) versus CLIPScore on random samples from the challenging Winoground [65] benchmark, containing real-world text prompts covering diverse compositional reasoning skills (which are carefully defined and labelled, as detailed in Appendix A). VQAScore performs well across basic compositions (attribute/scene/relation) as well as advanced compositions that require higher-order reasoning, e.g., counting attribute-object pairs and reasoning logically over negation and universality statements. Quantitative performance per skill can be found in Table 3.

models. GenAI-Bench exceeds the diversity and difficulty of prior benchmarks such as PartiPrompt [23, 28, 77]. We refer readers to [33] for further analysis on GenAI-Bench.

Extending to text-to-video/3D evaluation. Finally, we conduct preliminary experiments on video-text and 3D-text alignment benchmarks [44,72] by simply averaging the VQAScore across sampled frames or rendered views. VQAScore significantly surpasses popular methods such as CLIPScore [17], PickScore [28], and SOTA divide-and-conquer approaches that make use of GPT4-Vision [70].

Contribution summary.

1. We propose **VQAScore**, a simple metric that outperforms prior art without making use of expensive human feedback or proprietary models such as ChatGPT and GPT4-Vision.
2. VQAScore based on our proposed **CLIP-FlanT5** model achieves the state-of-the-art in vision-language alignment, offering a strong alternative to CLIPScore. We open-source a pip-installable API at link to run VQAScore for image/video/3D evaluation using one-line of Python code.
3. We present **GenAI-Bench**, a comprehensive benchmark with 1,600 compositional prompts to evaluate text-to-visual generation, surpassing the size and difficulty of existing benchmarks. Additionally, we provide over 15,000 human ratings (expanded to 80,000 in [33]) to support research on vision-language alignment metrics. Our dataset is available at link.

2 Related Works

Automated text-to-visual evaluation. Perceptual metrics like Inception Score (IS) [59], Fréchet Inception Distance (FID) [18] and Learned Perceptual Image Patch Similarity (LPIPS) [80] use pre-trained networks to assess the quality of generated imagery. However, these metrics rely on reference images and do not generalize to vision-language alignment. Recent text-to-visual systems [1–3, 13, 14, 19, 26, 27, 30–32, 35, 47, 57, 58, 60, 71] mostly report CLIPScore [17], which measures (cosine) similarity of the embedded image and text prompt. However, CLIP cannot reliably process compositional text prompts [25, 39, 65, 78]. Recent work further proposes three types of alignment metrics: **(1) Human-feedback approach.** ImageReward [74], PickScore [28], and HPSv2 [73] finetune VLMs like CLIP and BLIP on large-scale human ratings collected on generated images. **(2) GPT4-Vision-based approach.** VIEScore [29] and GPT4-Eval [81] carefully design a set of prompts for the proprietary GPT4-Vision [49] to output an image-text alignment score. **(3) Divide-and-conquer approach.** This popular line of methods [7, 23, 45, 61, 70] use LLMs such as ChatGPT to decompose texts into simpler components for analysis. A notable technique within this framework is Question Generation and Answering (QG/A), exemplified by TIFA [21], VQ2 [76], and Davidsonian [6]. For example, TIFA decomposes a text prompt into several simpler QA pairs and then outputs an alignment score as the accuracy of the answers generated by a VQA model.

Visio-linguistic compositional reasoning. Recent neuro-symbolic methods like visual programming [16,22,63] also use LLMs like ChatGPT to decompose complex visual tasks (described in natural language) into modular components. For instance, VPEval [7] applies visual programming to compute image-text alignment, using ChatGPT to invoke expert VLMs like image captioning [37] and open-vocabulary detection [43] models to examine fine-grained visual details. While visual programming achieves decent performance on classic benchmarks like GQA [24] and NLVR [62], we find that they rely heavily on hand-crafted in-context prompts (e.g., exemplar programs) and struggle on more challenging compositional tasks like Winoground [65]. Lastly, our VQAScore can be viewed as an extension of VisualGPTScore [39], which uses captioning models [37] to calculate the generative likelihood of $P(\text{text}|\text{image})$.

3 Image-Text Alignment Using VQAScore

This section describes how we compute VQAScore for image-text alignment, and introduces our CLIP-FlanT5 model that achieves the state-of-the-art.

Image-text alignment. Given an image $\mathbf{i}$ and a text $\mathbf{t}$, we aim to compute an alignment score $S(\mathbf{i}, \mathbf{t}) \in \mathbb{R}$, where higher scores reflect greater image-text similarity. Ideally, a model-predicted alignment score should closely match human judgment. For example, given the text "the moon is *over* the cow", an image incorrectly showing the cow *above* the moon would likely receive a lower human rating. Figure 3 provides such examples from the challenging Winoground [65]

benchmark. However, this seemingly simple task challenges contrastive VLMs like CLIP [25, 39, 78], which fail to understand *compositional* text prompts involving relations, attributes, and logical reasoning. Instead, we propose using recent *generative* VLMs trained for visual-question-answering (VQA), which can reason compositionally by generating answers based on images and questions.

Computing VQAScore. We calculate the alignment score directly from a VQA model using a simple template that converts the text $\mathbf{t}$ to a question $\mathbf{q}(\mathbf{t})$:

$$\begin{aligned} \mathbf{t} &= \text{The moon is over the cow} \\ \mathbf{q}(\mathbf{t}) &= \text{Does this figure show "The moon is over the cow"?} \\ &\quad \text{Please answer yes or no.} \end{aligned}$$

Next, we compute the generative likelihood of "Yes" from the auto-regressive answer decoder of an off-the-shelf VQA model (see Figure 1-a):

$$\text{VQAScore}(\mathbf{i}, \mathbf{t}) := P(\text{"Yes"} | \mathbf{i}, \mathbf{q}(\mathbf{t})) \tag{2}$$

Improving VQAScore via CLIP-FlanT5. While Eq. (2) can be readily computed using open-source models like LLaVA-1.5 [40], we improve VQAScore by training an in-house VQA model that follows best practices in the literature. Specifically, we find that popular VQA models [40, 41] are typically derived from "decoder-only" LLM architectures like Llama-2 [66] that use a uni-directional (auto-regressive) attention mechanism, where each token is influenced only by its previous tokens, but not vice versa. However, literature in language modeling [8,64] suggests that bidirectional encoder-decoders (where all tokens can influence each other) outperform the uni-directional counterparts on reasoning-focused linguistic tasks [5]. We argue that the architectural choice of image-question encoder becomes even more critical for visio-linguistic reasoning. For example, the state-of-the-art LLaVA-1.5 [40] places image tokens (MLP-projected CLIP visual tokens) ahead of question tokens. This prevents question tokens from influencing the preceding image tokens, which contradicts how humans process visual information based on prompted tasks [20]. Although training a new bidirectional LLM from scratch is not feasible due to substantial computational costs, we can still improve VQAScore by replacing Llama-2 in LLaVA-1.5 with the state-of-the-art bidirectional encoder-decoder FlanT5 [8] (see Figure 1-b for a comparison). For a fair comparison, we adhere to the training recipe of LLaVA-1.5, including the use of the same CLIP visual encoder, a modest 665K mixture of public VQA datasets, and a two-stage finetuning procedure. Appendix D includes more training details.

4 Experimental Results

This section outlines the experimental setup and results, highlighting VQAScore's superior performance compared to baseline methods such as CLIPScore [17], TIFA [21], and PickScore [28].

Baseline methods. We compare VQAScore against five popular method types: (1) VLM-based metrics (CLIPScore [17] and BLIPv2Score [37]); (2) VLMs finetuned on human feedback (PickScore [28], ImageReward [74], and HPSv2 [73]); (3) visual programming methods (VisProg [16], ViperGPT [63], and VPEval [7]); (4) divide-and-conquer methods using VQA (TIFA [21], VQ2 [76], and Davidsonian [6]); (5) approaches using proprietary models like GPT4-Vision (GPT4V-Eval [81] and VIEScore [29]). Appendix E describes all methods in detail.

Evaluating VQAScore on compositional image-text matching. We begin with the two most challenging benchmarks, Winoground [65] and EqBen [68], where each test sample has two (image, text) pairs. These benchmarks evaluate image-text matching through binary retrieval tasks that identify the best caption (from the pair of candidates) for a given image, and vice versa. Importantly, the benchmark API requires algorithms to return a match score for each candidate (image, text) pair instead of a relative ranking. This means they can be readily repurposed to evaluate image-text alignment. Compared to existing alignment benchmarks [21], we find that these matching benchmarks (especially Winoground) include more challenging text prompts with compositional structure (inspiring our own benchmarking efforts in Section 5). For example, the prompt ``someone talks on the phone angrily while another person sits happily'' requires the model to differentiate between two people (entities) based on emotions (attributes) and actions (relations). Another prompt ``three white and two brown eggs'' requires the model to count attribute-object pairs. Figure 3 compares VQAScore and CLIPScore on random Winoground examples. **??** provides an in-depth analysis of the skills covered by these benchmarks.

VQA achieves SOTA on matching benchmarks. Table 1 shows that VQAScore sets a new state-of-the-art on both benchmarks. Compared to baselines (e.g., CLIPScore [17] and PickScore [28]) that perform at chance-level, our VQAScore achieves 2x to 5x higher scores. Our results using open-source VQA models (e.g., InstructBLIP [9] and LLaVA-1.5 [40]) can match the previous SOTA method VQ2 [76] that uses the closed-source PaLI-17B [4] model, which was trained on 40x more private data (over 20 billion images and texts). Crucially, VQAScore based on our in-house CLIP-FlanT5 model surpasses all prior art, including two recent methods [29, 81] that use the proprietary (and expensive) GPT4-Vision [49] to score image-text alignment . Moreover, our experiments show that visual programming methods, including VisProg [16], ViperGPT [63], and VPEval [7], fail at compositional image-text matching, despite utilizing ChatGPT with expert VLMs [37, 43]. To fairly compare with divide-and-conquer methods that also use VQA models, we evaluate VQAScore against them based on the same VQA architectures, as demonstrated below.

End-to-end VQAScore outperforms divide-and-conquer methods. For a fair comparison, we apply three popular open-source divide-and-conquer methods (TIFA [21], VQ2 [76], Davidsonian [6]) with the same VQA models used for VQAScore. These methods either carefully prompt ChatGPT or finetune open-source LLMs like Llama-2 to decompose texts into simpler question-answer pairs. However, we discover that they struggle with compositional texts. For example,

Table 1: VQAScore achieves SOTA performance on challenging image-text matching benchmarks that require advanced compositional reasoning. We thoroughly ablate our proposed VQAScore with popular recent approaches on Winoground [65] and EqBen [68]. We strictly adhere to the original evaluation protocols and report text/image/group scores, with higher scores indicating better performance. We describe these metrics in Appendix F. Note that our VQAScore (highlighted in green) even matches or outperforms proprietary models (highlighted in gray) that appear to be trained on much more data (such as PALI-17B [4] and GPT4-Vision [49]).

Method	Models	Winoground			EqBen		
		Text	Image	Group	Text	Image	Group
Baselines							
Random Chance	–	25.0	25.0	16.7	25.0	25.0	16.7
Human Evaluation	–	89.5	88.5	85.5	–	–	–
Based on vision-language models							
CLIPScore [54]	CLIP-L-14	27.8	11.5	7.8	35.0	35.0	25.0
BLIPv2Score [37]	BLIPv2	43.3	21.3	17.5	48.6	43.6	35.0
Finetuned on human feedback							
PickScore [28]	CLIP-H-14 (finetuned)	23.8	12.5	6.8	35.7	39.3	23.6
ImageReward [74]	BLIPv2 (finetuned)	42.8	15.3	12.8	37.9	36.4	26.4
HPSv2 [73]	CLIP-H-14 (finetuned)	11.5	7.8	4.0	27.9	26.4	17.1
Based on visual programming							
VisProg [16]	ChatGPT, ViLT, OWL-ViT	3.5	3.5	3.5	7.9	7.9	7.9
ViperGPT [63]	ChatGPT, CLIP, BLIP, GLIP	7.8	7.8	7.8	4.3	4.3	4.3
VPEval [7]	ChatGPT, BLIP, GroundDINO	12.8	11.0	6.3	34.3	25.7	21.4
Divide-and-conquer via VQA							
VQ2 [76]	FlanT5, LLaVA-1.5	14.0	27.3	10.0	22.9	40.7	20.0
Davidsonian [6]	ChatGPT, LLaVA-1.5	21.0	16.8	15.5	26.4	20.0	20.0
Based on proprietary models							
TIFA [21, 76]	Llama-2, PaLI-17B	19.0	12.5	11.3	–	–	–
VQ2 [76]	FlanT5, PaLI-17B	47.0	42.0	30.5	–	–	–
GPT4V-Eval [81]	GPT4-Vision	44.5	49.0	36.3	42.9	40.0	35.0
VIEScore [29]	GPT4-Vision	40.8	39.3	34.5	40.0	34.3	32.9
VQAScore (ours) using open-source VQA model							
VQAScore	InstructBLIP	44.5	42.8	28.5	49.3	58.6	38.6
VQAScore	LLaVA-1.5	45.5	41.3	29.8	42.9	60.0	35.0
VQAScore (ours) using our VQA model							
VQAScore	**CLIP-FlanT5 (Ours)**	**60.0**	**57.5**	**46.0**	**59.3**	**63.6**	**47.9**

given ‘‘`someone talks on the phone angrily while another person sits happily`’’, Davidsonian [6] asks nonsensical questions like “is the someone talking angrily?” and “is the someone talking on the phone?”. Similarly, VQ2 [76] asks silly questions like “who talks with angrily on the phone?” and expects an answer of “someone”. Additionally, we find it crucial to expose the answer likelihood [39, 69], which is less biased than generating multiple-choice answers as done by [6, 21]. For instance, LLaVA-1.5 [40] biases towards answering “Yes” to 80% of the questions should be answered with “No” on Winoground (with the questions generated by Davidsonian [6]). **??** presents more analysis of these methods. Table 2 confirms that our simpler VQAScore significantly outperforms the more complex divide-and-conquer methods regardless of the underlying VQA models.

VQAScore can more effectively handle compositional text prompts. For a detailed analysis, we tag each Winoground sample by its associated compositional reasoning skills. **??** describes the labeling policy and procedure. Table 3

Table 2: Comparing VQAScore against divide-and-conquer methods using the same VQA models. For a fair comparison, we apply both VQAScore and three open-source divide-and-conquer methods (TIFA [21], VQ2 [76], and Davidsonian [6]) to the same underlying VQA architectures (InstructBLIP, LLaVA-1.5, and our CLIP-FlanT5). These popular methods make use of large language models to decompose compositional text prompts into simpler question-answer pairs for analysis, e.g., Llama-2 for TIFA, FlanT5 for VQ2, and ChatGPT for Davidsonian. However, they still struggle on compositional text prompts and often generate nonsensical question-answer pairs (more analysis can be found in Appendix E). Our end-to-end VQAScore (highlighted in green) outperforms them all using a much simpler question-answer template.

VQA Model	Method	Winoground			EqBen		
		Text	Image	Group	Text	Image	Group
–	Random Chance	25.0	25.0	16.7	25.0	25.0	16.7
InstructBLIP-FlanT5-11B [9]	TIFA [21]	20.3	16.3	14.5	25.0	25.7	18.6
	VQ2 [76]	19.0	26.3	11.3	20.0	39.3	15.7
	Davidsonian [6]	18.3	15.3	14.0	22.1	17.9	15.7
	VQAScore (ours)	**44.5**	**42.8**	**28.5**	**49.3**	**58.6**	**38.6**
LLaVA-1.5-13B [41]	TIFA [21]	22.8	18.5	15.5	30.0	30.0	21.4
	VQ2 [76]	14.0	27.3	10.0	22.9	40.7	20.0
	Davidsonian [6]	21.0	16.8	15.5	26.4	20.0	20.0
	VQAScore (ours)	**45.5**	**41.3**	**29.8**	**42.9**	**60.0**	**35.0**
CLIP-FlanT5-11B (Ours)	TIFA [21]	26.5	19.3	16.0	28.6	23.6	18.6
	VQ2 [76]	19.8	30.3	14.0	25.7	47.1	22.1
	Davidsonian [6]	16.3	11.5	9.8	17.1	11.4	11.4
	VQAScore (ours)	**60.0**	**57.5**	**46.0**	**59.3**	**63.6**	**47.9**

shows that VQAScore based on our CLIP-FlanT5 model significantly surpasses CLIPScore by 5x in basic skills (e.g., attribute, scene, relation) and 10x in advanced skills (e.g., counting, comparison, differentiation, negation, universality). Though trained on the same VQA data, our CLIP-FlanT5 (based on the 11B FlanT5 model) consistently outperforms LLaVA-1.5 (based on the 13B Llama-2 model). We believe our model benefits from the bidirectional image-question encoding and strong language capabilities of FlanT5, which has been finetuned on over 400 complex QA datasets [8]. Appendix D demonstrates that VQAScore can be improved by scaling up the language model and finetuning on VQA data.

Evaluating VQAScore's agreement with human judgments. We now test VQAScore on five text-to-image evaluation benchmarks (TIFA160 [21], Pick-a-Pic [28], and DrawBench [58], EditBench [67], COCO-T2I [38]) to measure its correlation (or agreement) with human judgments of alignment. In these benchmarks, given a text prompt, humans rate each generated image on a 1-to-5 Likert scale or assign a binary match-or-not label. Additionally, we report on an image-to-text evaluation benchmark Flickr8K [17], where each caption is manually rated based on the image. We follow SeeTrue [76] to report AUROC on DrawBench, EditBench, and COCO-T2I. For TIFA160 and Flickr8K, we evaluate pairwise accuracy as advocated by Deutsch et al. [12] (EMNLP'23 outstanding paper), since the original Kendall metric cannot handle ties common in human ratings. We report other metrics (e.g., Pearson and Kendall) in Appendix F. Due to the excessive noisy labels and NSFW content in the original Pick-a-pic

Table 3: Fine-grained analysis on Winoground. We report group scores per skill category. Note that each sample can naturally incorporate multiple skills. For instance, "`a white dog is on a brown couch`" involves understanding both "attribute" and "spatial relation". Additionally, a more complex prompt like "`six people wear blue shirts and no people wear white shirts`" requires higher-order reasoning (e.g., "counting" and "negation") along with other basic skills. We detail the skill definitions in Appendix A. Notably, the "advanced" skills (e.g., logic and comparison) prove more difficult (indicated by lower overall scores) compared to the "basic" skills. Our CLIP-FlanT5-based VQAScore excels across all skills – 5x better than CLIPScore on "basic skills" and 10x better on "advanced skills".

Method	Attribute	Scene	Relation			Overall
			Spatial	Action	Part	
CLIPScore (ViT-L-14)	13.0	40.0	8.5	11.1	11.5	9.9
VQAScore (InstructBLIP)	52.2	70.0	41.4	**50.0**	50.0	48.1
VQAScore (LLaVA-1.5)	53.6	**80.0**	47.6	27.8	57.7	47.3
VQAScore (CLIP-FlanT5)	**59.4**	**80.0**	**57.3**	44.4	**69.2**	**57.2**

(a) Basic skills (excluding samples requiring advanced skills)

Method	Count	Differ	Compare	Logical		Overall
				Negate	Universal	
CLIPScore (ViT-L-14)	7.8	2.3	2.0	0.0	0.0	4.4
VQAScore (InstructBLIP)	37.3	11.6	22.4	40.0	0.0	20.4
VQAScore (LLaVA-1.5)	29.4	20.9	16.3	40.0	0.0	24.1
VQAScore (CLIP-FlanT5)	**54.9**	**44.2**	**49.0**	**60.0**	**73.3**	**51.1**

(b) Advanced skills (including samples requiring basic skills)

dataset [28], we manually filter its testset, resulting in a clean subset of 100 samples (each has one prompt and two images) for evaluating binary accuracy.

VQAScore shows superior correlation with human judgments. Table 4 shows that VQAScore sets a new SOTA across all text-to-image alignment benchmarks, outperforming methods that finetune using costly human feedback [28, 73, 74] or rely on proprietary models [29, 76]. Appendix F also shows that VQAScore achieves a new SOTA on the image-to-text alignment benchmark Flickr8K, outperforming methods like CIDEr and RefCLIPScore that require additional reference captions [17]. Lastly, we highlight that the text prompts in these benchmarks lack the advanced compositional structure of real-world prompts (e.g., Winoground [65]). This motivates us to develop a benchmark with more challenging and realistic text prompts, which we present in Section 5.

5 GenAI-Bench for Text-to-Visual Evaluation

In this section, we introduce **GenAI-Bench**, a more challenging benchmark with compositional text prompts to evaluate both (1) text-to-visual generation models and (2) vision-language alignment metrics. Below, we present a preliminary study on GenAI-Bench, with further analysis in [33, 34].

Collecting GenAI-Bench. Inspired by the compositional structure of real-world (user-written) prompts [48,65], GenAI-Bench gathers text prompts covering essential visio-linguistic compositional reasoning skills, especially advanced ones (e.g., comparison, counting, logic) that are not fully explored in previous benchmarks, e.g., PartiPrompt [77], DrawBench [58], and T2I-CompBench [23]. First, we collaborate with graphic designers who routinely use text-to-image tools like Midjourney [48] to compile a comprehensive set of skills by surveying recent benchmarks [23, 58, 77] and real-world prompts [48]. Next, we collect prompts from these designers and ensure the prompts are relevant for real-world usage

Table 4: VQAScore on image-text alignment benchmarks that score agreement with human judgments of alignment. We show AUROC for DrawBench, EditBench, and COCO-T2I; pairwise accuracy [12] for TIFA160; and binary accuracy for Pick-a-Pick, with higher scores indicating better performance for all metrics. VQAScore (with CLIP-FlanT5) outperforms all prior art across all benchmarks. We find texts in these alignment benchmarks to lack the compositional structure compared to user-written prompts in benchmarks like [65], motivating us to create GenAI-Bench.

Method	Models	DrawBench	EditBench	COCO-T2I	TIFA160	Pick-a-Pic
Based on vision-language models						
CLIPScore [17]	CLIP-L-14	49.1	60.6	63.7	54.1	76.0
BLIPv2Score [37]	BLIPv2	60.5	68.0	70.7	57.5	80.0
Finetuned on human feedback						
PickScore [28]	CLIP-H-14 (finetuned)	72.3	64.3	61.5	59.4	70.0
ImageReward [74]	BLIPv2 (finetuned)	70.4	70.3	77.0	67.3	75.0
HPSv2 [73]	CLIP-H-14 (finetuned)	63.1	64.1	60.3	55.2	69.0
Divide-and-conquer via VQA						
VQ2 [76]	FlanT5, LLaVA-1.5	52.8	52.8	47.7	48.7	73.0
Davidsonian [6]	ChatGPT, LLaVA-1.5	78.8	69.0	76.2	54.3	70.0
Based on proprietary models						
TIFA [21, 76]	Llama-2, PaLI-17B	73.4	67.8	72.0	–	–
VQ2 [76]	FlanT5, PaLI-17B	82.6	73.6	83.4	–	–
GPT4V-Eval [81]	GPT4-Vision	–	–	–	64.0	74.0
VIEScore [29]	GPT4-Vision	–	–	–	63.9	78.0
VQAScore (ours) using open-source VQA models						
VQAScore	InstructBLIP	82.6	75.7	83.0	70.1	83.0
VQAScore	LLaVA-1.5	82.2	70.6	79.4	66.4	76.0
VQAScore (ours) using our VQA model						
VQAScore	**CLIP-FlanT5 (Ours)**	**85.3**	**77.0**	**85.0**	**71.2**	**84.0**

and free from subjective or toxic content, e.g., malicious web users often craft prompts with NSFW content [28]. Appendix B discusses the issues we found in previous benchmarks [23, 28, 77]. Lastly, we carefully tag each prompt with *all* its associated visio-linguistic skills, in contrast to previous benchmarks that either release no tags [28, 44, 73] or limit them to one or two [23, 58, 77]. The final GenAI-Bench contains 1,600 text prompts with over 5,000 human-verified skill tags. Appendix A details the skill definitions.

GenAI-Bench challenges leading text-to-visual models. Figure 4-a shows that state-of-the-art image and video generative models, such as DALL-E 3 [1], Stable Diffusion (SD-XL) [56], Pika [53], and Gen2 [15], struggle with GenAI-Bench's compositional text prompts that require higher-order reasoning such as comparison, differentiation, counting, and logic. Figure 4-b compares the averaged VQAScore (based on CLIP-FlanT5) of six image and four video generative models. We compute VQAScore for video-text pairs by averaging across all video frames as described in Section 6. We separately analyze each model's performance on "basic" and "advanced" prompts. Our analysis reveals significant improvements in text-to-visual generation for "basic" prompts from 2022 to 2023; however, improvements are less pronounced for "advanced" prompts, reflected in lower VQAScores across models. Nonetheless, we find that models with stronger language capabilities generally perform better. For example, one of the best open-source models DeepFloyd-IF [11] uses strong text embeddings from the T5 language model [55] rather than CLIP's, which do not encode compositional structure [25]. Similarly, the best closed-source model DALL-E

3 [1] does not directly train on noisy web text captions but instead improves them using captioning models. Finally, we anticipate significant advancements in open-source and video-generative models (e.g., SD-XL [56] and Gen2 [15]), which currently lag behind their closed-source and image-generative counterparts.

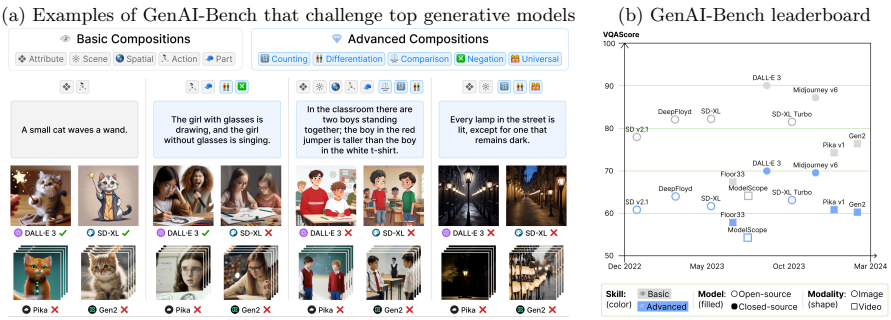

Fig. 4: GenAI-Bench. Figure (a) shows example prompts and associated skill tags from GenAI-Bench. The advanced compositional prompts of GenAI-Bench pose greater challenges to leading image and video generative models. **Figure (b)** presents the GenAI-Bench performance of 10 open/closed-source generative models. For each model, we separately show the averaged VQAScore for basic (in gray) and advanced (in blue) prompts. We find that (1) "advanced" prompts challenge all models more, (2) models that use stronger text embeddings or captions (e.g., DALL-E 3 [1] and DeepFloyd [11]) achieve the best results, (3) open-source and video generative models [15, 56] still lag behind their closed-source and image counterparts [1, 48], indicating potential for further improvement. **??** confirms that VQAScore agrees with collected human ratings.

VQAScore agrees with human judgments on GenAI-Bench. We hire three annotators to rate the image-text (or video-text) pairs on a 1-5 Likert scale, following the annotation protocol of [50]. In this work, we report on a subset of 527 prompts and collect 15,810 ratings across the ten generative models, significantly exceeding the scale of human evaluation in previous work [21, 58]. We extend our analysis to all 1,600 prompts in [33]. Appendix B confirms that VQAScore achieves the state-of-the-art correlation with human judgments on GenAI-Bench. We release all human ratings to support the development of alignment metrics.

6 Extending VQAScore to Video and 3D

We now use VQAScore to evaluate the alignment of text-to-video/3D models.

Video and 3D alignment benchmarks. For video-text alignment, we use the EvalCrafter [44] benchmark with 1-5 human Likert scales collected by [70]. For 3D-text alignment, we adopt the StanfordT23D benchmark [72], which released 3D assets but not human ratings. As such, we collect over 1,000 human 1-5 Likert scales on six text-to-3D models. We report Pairwise accuracy [12], Pearson, and Kendall on both benchmarks.

VQAScore achieves SOTA on video/3D-text alignment. To compute VQAScore using VQA models trained solely on images, we uniformly sample video frames across time and 2D views from 3D assets across camera angles. Table 5-a shows that our VQAScore surpasses the divide-and-conquer approach T2VScore-A [70] based on GPT4-Vision. Table 5-b shows that VQAScore exceeds popular text-to-3D metrics [72] such as CLIPScore [17] and PickScore [28]. In Appendix F, we show it is possible to achieve near-optimal performance using as few as 4 video frames (or 9 views), in contrast to the 36 video frames (or 120 views) provided by the original benchmarks.

Table 5: Evaluating VQAScore on text-to-video/3D benchmarks. We uniformly sample frames from videos and rendered views from 3D assets to calculate the average VQAScore (and other metrics). We report Pairwise accuracy, Pearson, and Kendall, with higher scores indicating better performance for all metrics. VQAScore surpasses popular video/3D metrics like CLIPScore [17], PickScore [28], and methods based on the proprietary GPT4-Vision [70] on both benchmarks.

Method	Pairwise Acc [12]	Old Metrics	
		Pearson	Kendall
Baselines reported in [70]			
CLIPScore	59.9	34.3	23.6
X-CLIPScore	56.9	25.7	17.5
BLIP-BLEU	53.0	15.2	10.4
T2VScore-A reported in [70]			
Otter-Video	–	18.1	13.4
Video-LLaMA	–	28.8	20.6
mPLUG-OWL2-Video	–	39.4	28.5
mPLUG-OWL2-Image	–	35.8	25.7
InstructBLIP	–	34.2	24.6
T2VScore-A w/ GPT4-V [70]			
GPT4-Vision	61.4	48.6	36.0
VQAScore w/ open-source models			
InstructBLIP	65.8	46.5	35.8
LLaVA-1.5	63.7	44.9	31.4
VQAScore w/ our model			
CLIP-FlanT5 (Ours)	**66.5**	**49.1**	**37.1**

(a) Text-to-video benchmark (T2VScore [70])

Method	Pairwise Acc [12]	Old Metrics	
		Pearson	Kendall
Baselines			
CLIPScore [17]	61.0	48.1	32.6
BLIPv2Score [37]	56.6	34.3	23.4
Finetuned on human feedback			
ImageReward [74]	66.3	57.1	43.8
PickScore [28]	60.1	41.3	30.8
HPSv2 [73]	55.9	31.5	21.9
VQAScore w/ open-source models			
InstructBLIP	68.0	59.5	47.5
LLaVA-1.5	64.9	55.8	40.8
VQAScore w/ our model			
CLIP-FlanT5 (Ours)	**68.6**	**64.3**	**48.7**

(b) Text-to-3D benchmark (StanfordT23D [72])

7 Conclusion

Limitations and future work. While VQAScore excels in vision-language alignment, it does not evaluate other critical aspects of generative models [32, 52], such as toxicity, bias, aesthetics, video motion, and 3D physics. We posit that VQAScore can evaluate these aspects if it were finetuned on relevant data.

Summary. We introduce VQAScore, a simple method surpassing current alignment metrics in evaluating text-to-image/video/3D models. VQAScore based on our CLIP-FlanT5 model offers a strong alternative to CLIPScore, especially on compositional text prompts. We introduce a more challenging GenAI-Bench to evaluate both text-to-visual generative models and automated alignment metrics. We hope our novel metric and benchmark will advance the scientific evaluation of generative models.

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 Harman Singh, 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.

References

1. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. https://cdn.openai. com/papers/dall-e-3.pdf (2023)
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
3. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
4. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
5. Chia, Y.K., Hong, P., Bing, L., Poria, S.: Instructeval: Towards holistic evaluation of instruction-tuned large language models. arXiv preprint arXiv:2306.04757 (2023)
6. Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. arXiv preprint arXiv:2310.18235 (2023)
7. Cho, J., Zala, A., Bansal, M.: Visual programming for text-to-image generation and evaluation. arXiv preprint arXiv:2305.15328 (2023)
8. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
9. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
10. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
11. Deepfloyd IF. `https://github.com/deep-floyd/IF` (2024)
12. Deutsch, D., Foster, G., Freitag, M.: 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. pp. 12914–12929 (2023)
13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
14. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022)

15. Gen2. https://research.runwayml.com/gen2 (2024)
16. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14953–14962 (2023)
17. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
19. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
20. Hochstein, S., Ahissar, M.: View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron **36**(5), 791–804 (2002)
21. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)
22. Hu, Y., Stretcu, O., Lu, C.T., Viswanathan, K., Hata, K., Luo, E., Krishna, R., Fuxman, A.: Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. arXiv preprint arXiv:2312.03052 (2023)
23. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. arXiv preprint arXiv:2307.06350 (2023)
24. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
25. Kamath, A., Hessel, J., Chang, K.W.: Text encoders bottleneck compositionality in contrastive vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 4933–4944 (2023)
26. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
27. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
28. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation (2023)
29. Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: Viescore: Towards explainable metrics for conditional image synthesis evaluation (2023)
30. Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., Chen, W.: Imagenhub: Standardizing the evaluation of conditional image generation models. arXiv preprint arXiv:2310.01596 (2023)
31. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
32. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H.B., Bellagente, M., et al.: Holistic evaluation of text-to-image models. arXiv preprint arXiv:2311.04287 (2023)

33. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: Evaluating and improving compositional text-to-visual generation. In: The First Workshop on the Evaluation of Generative Foundation Models at CVPR (2024)
34. Li, B., Lin, Z., Pathak, D., Li, J.E., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: GenAI-bench: A holistic benchmark for compositional text-to-visual generation. In: Synthetic Data for Computer Vision Workshop @ CVPR 2024 (2024), https://openreview.net/forum?id=hJm7qnW3ym
35. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench-2: Benchmarking multimodal large language models. arXiv preprint arXiv:2311.17092 (2023)
36. Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv preprint arXiv:2305.14720 (2023)
37. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
39. Lin, Z., Chen, X., Pathak, D., Zhang, P., Ramanan, D.: Revisiting the role of language priors in vision-language models. arXiv preprint arXiv:2306.01879 (2024)
40. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
41. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
42. Liu, S., Lin, Z., Yu, S., Lee, R., Ling, T., Pathak, D., Ramanan, D.: Language models as black-box optimizers for vision-language models. arXiv preprint arXiv:2309.05950 (2024)
43. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
44. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440 (2023)
45. Lu, Y., Yang, X., Li, X., Wang, X.E., Wang, W.Y.: Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. arXiv preprint arXiv:2305.11116 (2023)
46. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023)
47. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
48. Midjourney. https://www.midjourney.com (2024)
49. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
50. Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., Satoh, S.: Toward verifiable and reproducible human evaluation for text-to-image

generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14277–14286 (2023)
51. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
52. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails of vision-language models. arXiv preprint arXiv:2401.12425 (2024)
53. Pika. `https://www.pika.art/` (2024)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
55. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
57. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
58. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
59. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
60. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
61. Singh, J., Zheng, L.: Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. arXiv preprint arXiv:2307.04749 (2023)
62. Suhr, A., Lewis, M., Yeh, J., Artzi, Y.: A corpus of natural language for visual reasoning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 217–223 (2017)
63. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. arXiv preprint arXiv:2303.08128 (2023)
64. Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Bahri, D., Schuster, T., Zheng, H.S., Houlsby, N., Metzler, D.: Unifying language learning paradigms. arXiv preprint arXiv:2205.05131 (2022)
65. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)

66. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
67. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., 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. pp. 18359–18369 (2023)
68. Wang, T., Lin, K., Li, L., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Equivariant similarity for vision-language foundation models. arXiv preprint arXiv:2303.14465 (2023)
69. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)
70. Wu, J.Z., Fang, G., Wu, H., Wang, X., Ge, Y., Cun, X., Zhang, D.J., Liu, J.W., Gu, Y., Zhao, R., et al.: Towards a better metric for text-to-video generation. arXiv preprint arXiv:2401.07781 (2024)
71. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
72. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., Wetzstein, G.: Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. arXiv preprint arXiv:2401.04092 (2024)
73. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
74. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977 (2023)
75. Yang, Z., Wang, J., Li, L., Lin, K., Lin, C.C., Liu, Z., Wang, L.: Idea2img: Iterative self-refinement with gpt-4v (ision) for automatic image design and generation. arXiv preprint arXiv:2310.08541 (2023)
76. Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., Szpektor, I.: What you see is what you read? improving text-image alignment evaluation. arXiv preprint arXiv:2305.10400 (2023)
77. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
78. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: 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)
79. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
80. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
81. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023)