GenAI-Bench: A Holistic Benchmark for Compositional Text-to-Visual Generation

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

Abstract

Text-to-visual models can now generate photo-realistic images and videos that accurately depict objects and scenes. Still, they struggle with compositions of attributes, relationships, and higher-order reasoning such as counting, comparison, and logic. Towards this end, we introduce **GenAI-Bench** *to evaluate compositional text-to-visual generation through 1,600 high-quality prompts collected from professional designers, surpassing the difficulty and diversity of existing benchmarks like PartiPrompt and T2I-CompBench. Our human and automated evaluations on GenAI-Bench reveal that state-of-the-art models like DALL-E 3, StableDiffusion, and Gen2 often fail to parse user prompts requiring advanced compositional reasoning. Finally, we release over 24,000 human ratings on synthetic images and videos produced by ten leading generative models (with the numbers still growing) to support the development of automated text-to-visual evaluation metrics.*

1. Introduction

State-of-the-art text-to-visual models like Stable Diffusion [42], DALL-E 3 [2], and Sora [48] generate images and videos of exceptional quality. Due to their rapid advancement, traditional evaluation metrics and benchmarks (e.g., FID scores on the COCO dataset [18, 28]) are becoming insufficient [37]. For instance, in practical applications [36, 42], users often seek fine-grained control [3, 62] using *compositional* text prompts [33, 50] that involve attribute bindings, object relationships, and logical reasoning, among other visio-linguistic reasoning skills (Figure 1).

GenAI-Bench. We observe that existing benchmarks [19, 24, 34] such as PartiPrompt [60] and T2I-CompBench [20] do not fully capture the compositional structure of real-world user prompts. To remedy this, we identify a set of crucial skills for compositional text-to-visual generation, covering both basic (object, scene, attribute, relation) and advanced (counting, comparison, differentiation, logic) aspects. Next, we collect 1,600 diverse prompts from graphic designers who regularly use text-to-visual tools [36] for work. This approach ensures the quality and relevance of our prompts by excluding subjective and potentially toxic content crafted by malicious web users [24].

Human and automated evaluations. GenAI-Bench has collected over 24,000 human ratings for synthetically generated images and videos from ten leading models like DALL-E 3 [2], Midjourney v6 [36], Gen2 [15], and Pika [38]. Our preliminary study reveals that while these models can handle basic compositions (e.g., attributes and relations), they still struggle with higher-order reasoning like negation and comparison. Additionally, these human ratings enable us to benchmark automated metrics (e.g., CLIPScore [17]) that measure the alignment between an image and a text prompt. Specifically, we show that a simple end-to-end metric, **VQAScore** [30], derived from multimodal large language models (LLMs) [8, 31] trained for VQA, significantly outperforms CLIPScore and other carefully engineered metrics finetuned on human feedback (e.g., PickScore [24]) and question-generation-and-answering techniques (e.g., Davidsonian [5]). We will release the human ratings to support future benchmarking of automated evaluation metrics.

Contribution summary.

1. We present **GenAI-Bench**, a holistic benchmark with 1,600 quality prompts for compositional text-to-visual generation, surpassing the diversity and difficulty of previous benchmarks.
2. GenAI-Bench provides over 24,000 human ratings (with the number still growing) on synthetic images and videos to further research on automatic evaluation metrics for generative models.

2. Related Works

Text-to-visual benchmarks. Early benchmarks rely on captions from existing datasets like COCO [6, 19, 28,

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

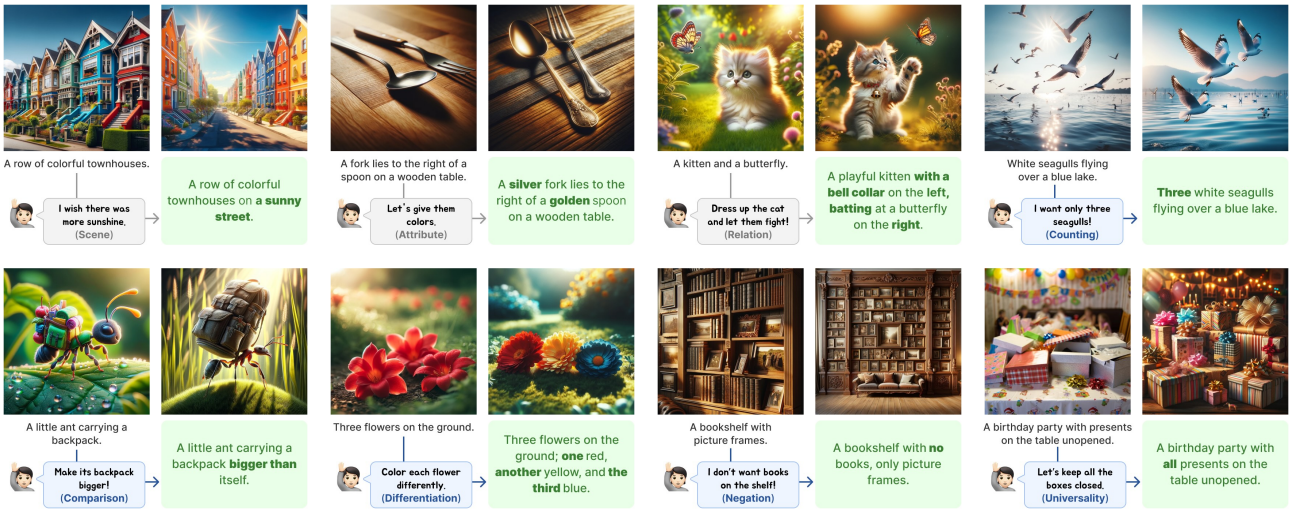


Figure 1. **Compositional text prompts of our GenAI-Bench (highlighted in green)** reflect how real-world users seek precise control in text-to-visual generation. For example, users might add details by specifying compositions of basic visual entities and properties (highlighted in gray), such as scenes, 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). We detail these skills and provide additional examples in the Appendix. Compared to previous benchmarks [20, 24, 44] like PartiPrompt [60], Table 1 shows that GenAI-Bench more comprehensively covers these essential aspects of compositional text-to-visual generation.

41], focusing on generating simple objects, attributes, and scenes. Other benchmarks, such as HPDv2 [57] and Pick-a-pic [24], primarily evaluate image quality (aesthetic) using simpler text prompts. Recently, DrawBench [44], PartiPrompt [60], and T2I-CompBench [20] 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 Inception Score (IS) [45], Fréchet Inception Distance (FID) [18] and Learned Perceptual Image Patch Similarity (LPIPS) [63] use pre-trained networks to assess the quality of generated imagery using reference images. For evaluate vision-language alignment (or faithfulness [9, 19]), recent studies [4, 13, 14, 23, 25, 35, 43, 46, 54] report CLIPScore [17], which measures (cosine) similarity of the embedded image and text prompt. However, CLIP cannot reliably process compositional text prompts [22, 29, 50, 61] due to its "bag-of-words" embeddings. Recent methods like ImageReward [58], PickScore [24], and HPSv2 [57] further leverage human-feedback to improve models like CLIP by finetuning on large-scale human ratings. Another popular line of works [7, 19, 20, 47, 55] uses LLMs like ChatGPT to decompose texts into simpler components for analysis, e.g., via question generation and answering (QG/A). For example, the Davidsonian method [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. [29, 30] show that such methods still face challenges with compositional prompts. Instead, they introduce an end-to-end metric called VQAScore: for a given image, it calculates the likelihood of a "Yes" answer to a simple question like "Does this figure show {text}?" VQAScore can be interpreted as the probability that the VQA model views the image as accurately reflecting the text, and it demonstrates a significantly stronger agreement with human judgment.

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 [20, 44, 60] 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). Figure 1 presents examples of these skills in GenAI-Bench. Table 1 shows that GenAI-Bench uniquely covers all these essential aspects. We provide definitions and more examples in Appendix A.

GenAI-Bench. We collect 1,600 prompts from designers who routinely use text-to-image tools [36]. 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 [20] can include subjective (e.g., non-visual) phrases like "`a natural symbol of rebirth and renewal`". Similarly, Pick-a-pic [24] 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 B. Lastly, we tag each prompt with *all* its evaluated aspects of compositional reasoning, in contrast to previous benchmarks that either release no tags [24, 34, 57] or limit them to one or two [20, 44, 60]. 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.

4. Evaluating Generative Models and Metrics

This section presents human and automated evaluations using GenAI-Bench for ten leading image and video generative models.

Human evaluation. We evaluate six text-to-image models: Stable Diffusion [42] (SD v2.1, SD-XL, SD-XL Turbo), DeepFloyd-IF [10], Midjourney v6 [36], DALL-E 3 [2]; along with four text-to-video models: ModelScope [51], Floor33 [12], Pika v1 [38], Gen2 [15]. Next, we collect 1-5 Likert scale human ratings for image-text or video-text alignment using the recommended annotation protocol of [37]:

> 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.

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 [19].

Automated evaluation. We use recent multimodal LLMs [8, 31] trained for VQA to compute the alignment score. Given an `image` and `text`, we calculate the **VQAScore** [30] defined as the probability of a "Yes" answer to a simple `question` like "Does this figure show '`{text}`'? Please answer yes or no.":

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

We implement VQAScore on an in-house VQA model **CLIP-FlanT5** (which we will release) trained on the 665K public VQA data from LLaVA-1.5 [31]. We attach implementation details in the Appendix. Despite its simplicity, Table 2 shows that VQAScore achieves the best correlation with human ratings on GenAI-Bench, outperforming previous methods including CLIPScore [17], models trained with extensive human feedback [24, 57, 58], and QG/A methods that use the same CLIP-FlanT5 VQA model [5, 59]. In Appendix, we also show that VQAScore achieves the state-of-the-art performance on more alignment benchmarks such as TIFA160 [19] and Winoground [50].

GenAI-Bench challenges leading text-to-visual models. Figure 2-a shows that state-of-the-art image and video generative models still struggle with GenAI-Bench's compositional text prompts. Figure 2-b compares the averaged VQAScore (based on CLIP-FlanT5) of the ten image and video generative models. We compute VQAScore for video-text pairs by averaging across all video frames following prior work [46]. 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 scores 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 [10] uses strong text embeddings from the T5 language model [40] rather than CLIP's, which do not encode compositional structure [22]. Similarly, the best closed-source model DALL-E 3 [2] 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 [42] and Gen2 [15]), which currently lag behind their closed-source and image-generative counterparts. We include per-skill human and VQAScore results in the Appendix.

5. Conclusion

Limitations. GenAI-Bench currently does not evaluate other aspects of generative models [26, 34, 56], such as toxicity, bias, aesthetics, and video motion.

Summary. We introduce a more challenging GenAI-Bench to benchmark both compositional text-to-visual generation and automated evaluation metrics, in hope of advancing the scientific evaluation of generative models.

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) [60]	✓	✓	✓	✓	✓	✗	✗	✗	2 Tags	✗
DrawBench [44]	✓	✓	✓	✓	✗	✗	✗	✗	1 Tag	✗
EditBench [52]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
TIFAv1 [19]	✓	✓	✓	✓	✗	✗	✗	✗	All Tags	Images
Pick-a-pic [24]	✓	✓	✓	✓	✗	✗	✗	✗	✗	Images
T2I-CompBench [20]	✓	✓	✓	✓	✗	✗	✗	✗	1 Tag	Not Released
HPDv2 [57]	✓	✓	✓	✗	✗	✗	✗	✗	✗	Images
EvalCrafter [34]	✓	✓	✓	✓	✗	✗	✗	✗	✗	Videos
GenAI-Bench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	All Tags	Images & Videos

Table 2. **Evaluating the human correlation of automated metrics on GenAI-Bench.** We report Pairwise accuracy [11], Pearson, and Kendall, with higher scores indicating better performance for all metrics. Our VQAScore based on the in-house CLIP-FlanT5 model (detailed in the Appendix) achieves the strongest agreement with human ratings on images and videos, significantly surpassing popular metrics like CLIPScore [17], PickScore [24], and Davidsonian [5].

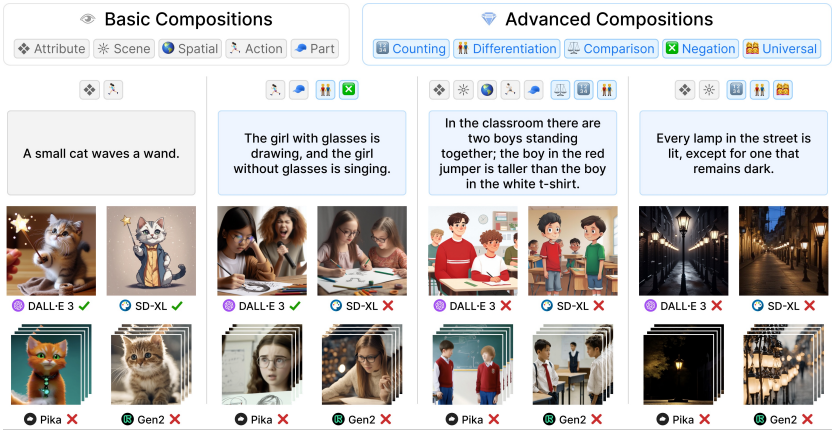
Method	Pairwise	Pearson	Kendall
CLIPScore [17]	51.9	19.3	13.5
ImageReward [58]	57.4	36.3	25.2
PickScore [24]	57.7	36.6	25.9
HPSv2 [57]	50.1	15.1	10.3
VQ2 [59]	52.5	16.2	14.8
Davidsonian [5]	54.2	32.5	23.1
VQAScore (Ours)	**63.3**	**46.0**	**38.0**

(a) GenAI-Bench (Image)

Method	Pairwise	Pearson	Kendall
CLIPScore [17]	54.2	26.5	18.5
ImageReward [58]	60.4	43.6	32.0
PickScore [24]	56.2	32.5	23.2
HPSv2 [57]	50.6	17.5	12.1
VQ2 [59]	52.8	18.0	15.5
Davidsonian [5]	55.9	32.3	23.5
VQAScore (Ours)	**64.4**	**53.3**	**39.9**

(b) GenAI-Bench (Video)

(a) Examples of GenAI-Bench that challenge top generative models



(b) GenAI-Bench leaderboard

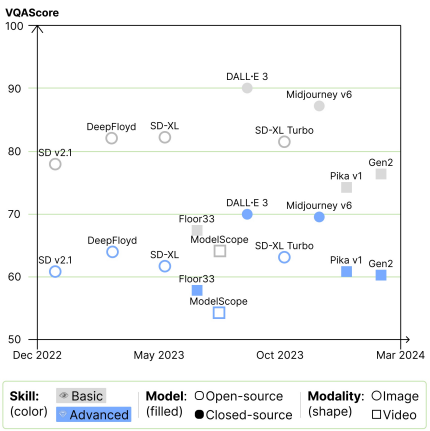


Figure 2. **GenAI-Bench. Figure (a)** shows example prompts and associated skill tags from GenAI-Bench. The advanced compositional prompts of GenAI-Bench pose greater challenges to leading image and video generative models. **Figure (b)** presents the GenAI-Bench performance of 10 open/closed-source 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 [2] and DeepFloyd [10]) outperform others (e.g, SD-XL [42]), (3) open-source and video generative models [15, 42] still lag behind their closed-source and image counterparts [2, 36], indicating potential for further improvement.

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. 10

[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, 8, 10, 12

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[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. 2

[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, 3, 4, 10, 12, 13

[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. 1, 8, 12

[7] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. *arXiv preprint arXiv:2305.15328*, 2023. 2

[8] 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. 1, 3, 10

[9] 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. 2

[10] Deepfloyd IF. Deepfloyd IF. https://github.com/deep-floyd/IF, 2024. 3, 4, 8

[11] 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. 4, 13

[12] Floor33. Floor33. https://www.morphstudio.com/, 2023. 3, 8, 10

[13] 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. 2

[14] 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. 2

[15] Gen2. Gen2. https://research.runwayml.com/gen2, 2024. 1, 3, 4, 8, 10

[16] 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. 11

[17] 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, 2, 3, 4, 12, 13

[18] 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, 2

[19] 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, 8, 10, 12

[20] 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, 8, 12

[21] 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. 11

[22] 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, 12

[23] 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. 2

[24] 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, 12, 13

[25] 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. 2

[26] 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. 3

[27] 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. 10, 12

[28] 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, 12

[29] 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. 2, 12

[30] 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, 3

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3, 10

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 10

[33] 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

[34] 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

[35] 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. 2

[36] Midjourney. Midjourney. https://www.midjourney.com, 2024. 1, 3, 4, 8

[37] 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, 3

[38] Pika. Pika. https://www.pika.art/, 2024. 1, 3, 8, 10

[39] 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. 10

[40] 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. 3

[41] 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. 2

[42] 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, 8

[43] 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. 2

[44] 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, 8, 12

[45] 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. 2

[46] 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. 2, 3

[47] 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. 2

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

[49] 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. 12

[50] 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, 3, 8, 12

[51] 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. 3, 8

[52] 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. 4

[53] 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. 12

[54] 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. 2

[55] 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. 2, 12

[56] 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. 3

[57] 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. 2, 3, 4, 12, 13

[58] 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, 3, 4, 12, 13

[59] 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. 3, 4, 12, 13

[60] 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, 8, 12

[61] 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. 2, 12

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[63] 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. 2

A Holistic Benchmark for Compositional Text-to-Visual Generation

Supplementary Material

Outline

This document supplements the main paper with benchmark and method details. Below is the outline:

- **Section A** details GenAI-Bench's evaluated aspects.
- **Section B** describes how we collect GenAI-Bench.
- **Section C** describes how we compute VQAScore.
- **Section D** describes our in-house VQA model (CLIP-FlanT5).
- **Section E** discusses other baseline methods.
- **Section F** discusses other alignment benchmarks.

A. Evaluated Aspects of GenAI-Bench

This section details the evaluated aspects of GenAI-Bench.

Skill definitions. Most literature on text-to-visual generation [6, 19, 20, 44, 60] primarily focuses on generating *basic* objects, attributes, relations, and scenes. While these "basic" visual compositions still pose challenges, real-world user prompts often introduce greater complexity. Such prompts require higher-order reasoning beyond basic compositions, including comparison, differentiation, counting, and logic. For example, while existing benchmarks focus only on counting objects [19, 60], real-world prompts often require counting attribute-object pairs or even object-relation-object triplets, like "`one person wearing a white shirt and the other five wearing blue shirts`". To this end, after thoroughly reviewing relevant literature [20, 36, 50, 60], we define a set of compositional reasoning skills common in real-world prompts, categorizing them into "basic" and "advanced", where the latter can build upon the former. For logical reasoning, we consider "negation" and "universality", which are the two most common types of logic we see in real-world prompts. We provide detailed definitions for "basic" skills in Table 3 and "advanced" skills in Table 4.

Comparing skills across benchmarks. We find the skill categorization in benchmarks like PartiPrompt [60] to be ambiguous or even confusing. For example, PartiPrompt introduces two categories "*complex*" and "*fine-grained detail*". The former refers to "*...fine-grained, interacting details or relationships between multiple participants*", while the latter refers to "*...attributes or actions of entities or objects in a scene*". Upon closer examination, the categorization of spatial, action, and part relations into these categories appears arbitrary. To address this, we attempt to compare the skill coverage across all benchmarks by our unified set of skills. For benchmarks (PartiPrompt/T2I-CompBench) with pre-defined skill categories, we map their skills to our definitions. For benchmarks (TIFAv1/Pick-a-pic/DrawBench/EditBench/HPDv2/EvalCrafter) without a comprehensive skill set, we manually annotate a random subset of samples. Finally, we calculate the skill proportions in each benchmark, identifying skills that constitute more than 2% as genuinely present.

B. GenAI-Bench

This section describes how we collect GenAI-Bench.

Details of GenAI-Bench. GenAI-Bench consists of 1,600 diverse prompts that cover advanced skills not addressed in previous benchmarks [20, 44, 60]. To source prompts relevant to real-world applications, we employ two graphic designers experienced in text-to-visual tools like Midjourney [36]. First, we introduce them to our skill definitions and examples. Then, we ask them to craft prompts for each skill, collaborating with ChatGPT to brainstorm prompt variants across diverse visual domains. Importantly, these designers ensure that the prompts are *objective*. This contrasts with T2I-CompBench [20], whose prompts are almost entirely auto-generated. For example, in T2I-CompBench's "*texture*" category, an overwhelming 40% of the 1000 programmatically-generated prompts use "metallic" as the attribute, which limits their diversity. Other T2I-CompBench's prompts generated by ChatGPT often contain subjective (non-visual) phrases. For instance, in the prompt "`the delicate, fluttering wings of the butterfly signaled the arrival of spring, a natural symbol of rebirth and renewal`", the "rebirth and renewal" can convey different meanings to different people. Similarly, in "`the soft, velvety texture of the rose petals felt luxurious against the fingertips, a romantic symbol of love and affection`", the "love and affection" is also open to diverse interpretations. Thus, we carefully guide the designers to avoid such prompts. Lastly, each prompt in GenAI-Bench is tagged with all its evaluated aspects. We streamline this process by using GPT4 for automatic tagging, providing it the skill definitions and in-context exemplars. Later, we manually verify and correct all tags for accuracy, resulting in over 5,000 human-verified tags.

Collecting human ratings. We evaluate six text-to-image models: Stable Diffusion [42] (SD v2.1, SD-XL, SD-XL Turbo), DeepFloyd-IF [10], Midjourney v6 [36], DALL-E 3 [2]; along with four text-to-video models: ModelScope [51], Floor33 [12], Pika v1 [38], Gen2 [15]. Due

Table 3. **Skill definitions and examples for basic compositions.**

Skill Type	Definition	Examples
	Basic Compositions	
`Object`	Basic entities within an image, such as person, animal, food, items, vehicles, or text symbols (e.g., "A", "1+1").	*a* **dog***, a* **cat** *and a* **chicken** *on a* **table***; a young* **man** *with a green* **bat** *and a blue* **ball***; a '***No Parking***' sign on a busy street.*
`Attribute`	Visual properties of entities, such as color, material, emotion, size, shape, age, gender, state, and so on.	*a* **silver** *spoon lies to the left of a* **golden** *fork on a* **wooden** *table; a* **green** *pumpkin is smiling* **happily***, a* **red** *pumpkin is sitting* **sadly**.
`Scene`	Backgrounds or settings of an image, such as weather, location, and style.	*A child making a sandcastle on a* **beach in a cloudy day***; a grand fountain surrounded by historic buildings in a* **town square**.
`Spatial Relation`	Physical arrangements of multiple entities relative to each other, e.g., on the right, on top, facing, towards, inside, outside, near, far, and so on.	*a bustling city street, a neon 'Open 24 Hours' sign glowing* **above** *a small diner; a teacher standing* **in front of** *a world map in a classroom; tea steams* **in** *a cup,* **next to** *a closed diary with a pen resting* **on** *its cover.*
`Action Relation`	Action interactions between entities, e.g., pushing, kissing, hugging, hitting, helping, and so on.	*a dog* **chasing** *a cat; a group of children* **playing** *on the beach; a boat* **glides** *across the ocean, dolphins* **leaping** *beside it and seagulls* **soaring** *overhead.*
`Part Relation`	Part-whole relationships between entities – one entity is a component of another, such as body part, clothing, and accessories.	*a pilot* **with aviator sunglasses***; a baker* **with a cherry pin on a polka dot apron***.; a young lady* **wearing a T-shirt** *puts* **her hand** *on a* **puppy's head**.

Table 4. **Skill definitions and examples for advanced compositions.**

Skill Type	Definition	Examples
	Advanced Compositions	
`Counting`	Determining the quantity, size, or volume of entities, e.g., objects, attribute-object pairs, and object-relation-object triplets.	**two** *cats playing with a* **single** *ball;* **five** *enthusiastic athletes and* **one** *tired coach;* **one** *pirate ship sailing through space, crewed by* **five** *robots;* **three** *pink peonies and* **four** *white daisies in a garden.*
`Differentiation`	Differentiating objects within a category by their attributes or relations, such as distinguishing between "old" and "young" people by age, or "the cat on top of the table" versus "the cat under the table" by their spatial relations.	**one cat** *is sleeping on the table and* **the other** *is playing under the table; there are two men in the living room,* **the taller one** *to the left of* **the shorter one***; a notebook lies open in the grass, with sketches on* **the left page** *and blank space* **on the right***; there are two shoes on the grass,* **the one without laces** *looks newer than* **the one with laces**.
`Comparison`	Comparing characteristics like number, attributes, area, or volume between entities.	*there are* **more** *people standing than sitting; between the two cups on the desk, the* **taller** *one holds* **more** *coffee than the* **shorter** *one, which is half-empty; a small child on a skateboard has* **messier** *hair than the person next to him; three little boys are sitting on the grass, and the boy in the middle looks the* **strongest**.
`Negation`	Specifying the absence or contradiction of elements, as indicated by "no", "not", or "without", e.g., entities not present or actions not taken.	*four elephants,* **no** *giraffes; six people wear white shirts and* **no** *people wear red shirts; a bookshelf with* **no** *books, only picture frames.; a person with short hair is crying while a person with long hair* **is not***; a smiling girl with short hair and* **no** *glasses.; a cute dog* **without** *a collar.*
`Universality`	Specifying when every member of a group shares a specific attribute or is involved in a common relation, indicated by words like "every", "all", "each", "both".	*in a room,* **all** *the chairs are occupied except one; a bustling kitchen where* **every** *chef is preparing a dish; in a square, several children are playing,* **each** *wearing a red T-shirt; a table laden with apples and bananas, where* **all** *the fruits are green; the little girl in the garden has roses in* **both** *hands.*

to the lack of APIs for Floor33 [12], Pika v1 [38], and Gen2 [15], we manually download videos from their websites. For image generative models, we generate images using all 1,600 GenAI-Bench prompts. We use a coreset of 800 prompts to collect videos for the four video models. The same 800 prompts are used to collect the ranking benchmark in the main paper. In total, we collect over 80,000 human ratings, greatly exceeding the scale of human annotations in previous work [5, 19], e.g., TIFA160 collected 2,400 ratings.

GenAI-Bench performance. We detail the performance of the ten image and video generative models across all skills in Table 5. Both humans and VQAScores rate DALL-E 3 [2] higher than the other models in nearly all skills, except for negation. In addition, prompts requiring "advanced" compositions are rated significantly lower by both humans and VQAScores, with negation being the most challenging skill. Lastly, current video models do not perform as well as image models, suggesting room for improvement.

C. Implementing VQAScore

In this section, we describe how we compute VQAScore.

Computing VQAScore as an auto-regressive product. Recall that VQAScore calculates the alignment score of an image **i** and text **t** directly from a VQA model. We first use a simple QA template to convert the text **t** to a question and an answer (denoted as **q**(**t**) and **a**(**t**)), for example:

$$\mathbf{q}(\mathbf{t}) = \text{Does this figure show "}\{\text{t}\}\text{"? Please answer yes or no.} \tag{2}$$

$$\mathbf{a}(\mathbf{t}) = \text{Yes} \tag{3}$$

We later demonstrate that such a straightforward question-answer pair is sufficient for good performance. In language modeling [1], a piece of text is pre-processed (or tokenized) into a token sequence, e.g., $\mathbf{a}(\mathbf{t}) = \{a_1, \cdots, a_m\}$. Although "`Yes`" usually counts as a single token, we include the EOS (end-of-sentence) token at the end of the text sequence for a simpler implementation. We find that the EOS token only marginally affects the VQAScore results. Next, the generative likelihood of the answer (conditioned on both the question and image) can be naturally factorized as an auto-regressive product [1]:

$$\text{VQAScore}(\mathbf{i}, \mathbf{t}) := P(\mathbf{a}(\mathbf{t})|\mathbf{i}, \mathbf{q}(\mathbf{t})) = \prod_{k=1}^{m} P(a_k|a_{<k}, \mathbf{i}, \mathbf{q}(\mathbf{t})) \tag{4}$$

The answer decoders of VQA models [8, 32] return back m softmax distributions corresponding to the m terms in the above expression. Computing VQAScore is more efficient than generating answer token-by-token. Since the entire sequence of tokens $\{a_k\}$ is already available as input for VQAScore, the above m terms can be efficiently computed in *parallel*. In contrast, answer generation as done by [5, 19] requires *sequential* token-by-token prediction, as token a_k must be generated before it can serve as input to generate the softmax distribution for the subsequent token a_{k+1}.

Algorithm 1: PyTorch-style pseudocode for VQAScore.

```
# tokenize():  text tokenizer that converts texts
 to a list of token indices
# vqa_model():  VQA model returns logits for
 predicted answer

def vqa_score(image, text):
    # Format the text into the below QA pair
    question = f"Does this figure show '{text}'?
     Please answer yes or no."
    answer = "Yes"

    # Tokenize the QA pair into tokens
    question_tokens = tokenize(question)
    answer_tokens = tokenize(answer)

    # Extract logits for predicted answer of shape
     [len(answer_tokens), vocab_size]
    # answer_tokens is a required input for
     auto-regressive decoding
    logits = vqa_model(image, question_tokens,
     answer_tokens)

    # labels must skip the first BOS
     (Begin-Of-Sentence) token
    labels = answer_tokens[1:]
    # logits must skip the last EOS
     (End-Of-Sentence) token
    logits = logits[:-1]

    # Compute the log likelihood of the answer
    log_likelihood =
     -torch.nn.CrossEntropyLoss()(logits, labels)
    # (Optional) Cancel the log to obtain P("Yes"
     | image, question)
    score = log_likelihood.exp()
    return score
```

Pseudocode of VQAScore. To better explain how VQAScore works, we attach the pseudocode in algorithm 1. We will release a pip-installable API to compute VQAScore using one-line of Python code.

D. Training CLIP-FlanT5

In this section, we detail the training procedure of CLIP-FlanT5.

Training CLIP-FlanT5. We adhere to the training recipe of the state-of-the-art LLaVA-1.5 [31]. We adopt the same (frozen) CLIP visual encoder (ViT-L-336) [39] and the 2-layer MLP projector for image tokenization. We also follow LLaVA-1.5's two-stage finetuning procedure and datasets. In stage-1 training, we finetune the MLP projector on 558K captioning data (LAION-CC-SBU with BLIP captions [27]). To accommodate FlanT5's encoder-decoder architecture, we adopt the split-text training method proposed in BLIPv2 [27]. This involves splitting a caption

Table 5. **Performance breakdown on GenAI-Bench.** We present the averaged human ratings and VQAScores (based on CLIP-FlanT5) for "basic" and "advanced" prompts. Human ratings use a 1-5 Likert scale, and VQAScore ranges from 0 to 1, with higher scores indicating better performance for both. Generally, both human ratings and VQAScores favor DALL-E 3 over other models, with DALL-E 3 preferred across almost all skills except for negation. We find that "advanced" prompts that require higher-order reasoning present significant challenges. For instance, the state-of-the-art DALL-E 3 receives a remarkable average human rating of 4.3 for "basic" prompts, indicating the images and prompts range from "*having a few minor discrepancies*" to "*matching exactly*". However, it scores only 3.4 for "advanced" prompts, suggesting "*several minor discrepancies*". In addition, video models receive significantly lower scores than image models. Overall, VQAScores closely match human ratings.

Method	Attribute	Scene	Relation			Avg
			Spatial	Action	Part	
Image models						
SD v2.1	3.3	3.3	3.0	3.2	3.1	3.2
SD-XL Turbo	3.7	3.7	3.4	3.5	3.5	3.6
SD-XL	3.8	3.7	3.4	3.7	3.6	3.6
DeepFloyd-IF	3.7	3.7	3.7	3.7	3.6	3.7
Midjourney v6	4.0	3.9	3.7	4.0	4.0	3.9
DALL-E 3	4.3	4.5	4.2	4.2	4.2	4.3
Video models						
ModelScope	3.1	3.1	2.8	3.0	3.1	3.0
Floor33	3.2	3.2	2.9	3.2	3.1	3.1
Pika v1	3.4	3.4	3.1	3.3	3.2	3.3
Gen2	3.6	3.7	3.4	3.6	3.6	3.6

(a) Human ratings on "basic" prompts

Method	Attribute	Scene	Relation			Avg
			Spatial	Action	Part	
Image models						
SD v2.1	0.80	0.81	0.76	0.77	0.79	0.79
SD-XL Turbo	0.83	0.83	0.80	0.81	0.84	0.83
SD-XL	0.86	0.86	0.82	0.83	0.89	0.84
Midjourney v6	0.89	0.89	0.87	0.87	0.91	0.87
DALL-E 3	0.91	0.91	0.91	0.89	0.91	0.90
Video models						
ModelScope	0.69	0.69	0.65	0.65	0.70	0.66
Floor33	0.70	0.71	0.64	0.66	0.67	0.67
Pika v1	0.78	0.80	0.74	0.72	0.76	0.75
Gen2	0.79	0.81	0.74	0.76	0.83	0.77

(b) VQAScores on "basic" prompts

Method	Count	Differ	Compare	Logical		Avg
				Negate	Universal	
Image models						
SD v2.1	2.7	2.4	2.5	2.7	2.9	2.8
SD-XL	2.8	2.6	2.5	2.7	3.2	2.8
SD-XL Turbo	2.8	2.5	2.6	2.8	3.2	2.9
DeepFloyd-IF	3.1	2.8	2.9	2.8	3.3	3.0
Midjourney v6	3.3	3.1	3.1	2.9	3.5	3.2
DALL-E 3	3.4	3.3	3.4	2.8	3.7	3.4
Video models						
ModelScope	2.4	2.4	2.2	2.6	2.8	2.5
Floor33	2.7	2.7	2.5	2.8	3.2	2.8
Pika v1	2.7	2.7	2.6	2.9	3.3	2.9
Gen2	2.8	2.7	2.6	2.9	3.3	2.9

(c) Human ratings on "advanced" prompts

Method	Count	Differ	Compare	Logical		Avg
				Negate	Universal	
Image models						
SD v2.1	0.67	0.67	0.66	0.55	0.59	0.62
SD-XL	0.71	0.71	0.72	0.53	0.62	0.64
SD-XL Turbo	0.70	0.69	0.71	0.55	0.61	0.65
DeepFloyd-IF	0.70	0.69	0.71	0.52	0.64	0.65
Midjourney v6	0.76	0.78	0.77	0.53	0.70	0.70
DALL-E 3	0.80	0.81	0.77	0.53	0.72	0.71
Video models						
ModelScope	0.58	0.61	0.57	0.52	0.52	0.55
Floor33	0.60	0.64	0.59	0.53	0.55	0.57
Pika v1	0.65	0.64	0.63	0.55	0.63	0.61
Gen2	0.69	0.69	0.64	0.54	0.58	0.62

(d) VQAScores on "advanced" prompts

into two parts at a random position, with the first part sent to the encoder and the second part to the decoder. In stage-2 training, we finetune both the MLP projector and the language model (FlanT5) on 665K mixture of public VQA datasets (e.g., VQAv2 [16] and GQA [21]). To efficiently train the encoder-decoder architecture, we convert all multi-turn VQA samples into single-turn, resulting in 3.4M image-question-answer pairs. We also retrain LLaVA-1.5 on the same single-turn VQA samples and observe the same VQAScore results. We borrow hyperparameters of LLaVA-1.5 (see Table 6), such as the learning rate schedule, optimizer, number of epochs, and weight decay. We use 8 A100 (80Gbs) GPUs to train all our models. Our largest CLIP-FlanT5-XXL (11B) takes 5 hours for the stage-1 and 80 hours for the stage-2. For stage-2 training, we adhere to the system (prefix) prompt of LLaVA-1.5 during training [1]:

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
> **USER:** image \n question **ASSISTANT:** answer

E. Details of Baseline Methods

In this section, we detail the implementation of the baseline methods. Note that Table 7 reports VQAScore perfor-

[1] By default, we also use the system prompt during inference. Interestingly, removing the system prompt ("A chat between a curious user ... answers to the user's questions") during inference does not affect CLIP-FlanT5 but will hurt LLaVA-1.5's performance.

Table 6. **Training hyperparameters for CLIP-FlanT5.**

Hyperparameter	Stage-1	Stage-2
dataset size	558K	665K
batch size	256	96
lr	1e-2	2e-5
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
epoch	1	
optimizer	AdamW	
DeepSpeed stage	2	3

mance on seven more benchmarks that measures correlation with human judgments.

CLIPScore and BLIPv2Score. To calculate CLIPScore, we use the same CLIP-L-336 model [17] of CLIP-FlanT5. To calculate BLIPv2Score, we use the ITM head of BLIPv2-vit-G [27]. For an in-depth analysis of how these discriminatively pre-trained VLMs behave as bag-of-words models, we refer readers to previous studies [22, 29, 50, 61].

Metrics finetuned on human feedback (PickScore/ImageReward/HPSv2). We use the official code and model checkpoints to calculate these metrics. Specifically, PickScore [24] and HPSv2 [57] finetune the CLIP-H model, and ImageReward [58] finetunes the BLIPv2 [27], using costly human feedback from either random web users or expert annotators. Our experiments on the Winoground and EqBen benchmarks (Table 7) show that these metrics perform no better than random chance, likely because the discriminative pre-trained VLMs bottleneck their performance due to bag-of-words encodings. In addition, their finetuning datasets may lack compositional texts. Finally, we observe that human annotations can be noisy or subjective, especially when these annotators are not well trained (e.g., random web users of the Pick-a-pic dataset [24]). We discuss these issues in Appendix F.

QG/A methods (VQ2/Davidsonian). We first note that these divide-and-conquer methods are the most popular in recent text-to-visual evaluation [2, 20, 49, 55]. VQ2 [59] uses a finetuned FlanT5 to generate free-form QA pairs and computes the average score of P(answer | image, question). Davidsonian uses a more sophisticated pipeline by prompting ChatGPT to generate yes-or-no QA pairs while avoiding inconsistent questions. For example, given the text "the moon is over the cow", if a VQA model already answers "No" to "Is there a cow?", it then skips the follow-up question "Is the moon over the cow?". However, these methods often generate nonsensical QA pairs, as shown in Table 8 on real-world user prompts from GenAI-Bench.

F. Details of Alignment Benchmarks

This section discusses other benchmarks.

TIFA160 [19]. TIFA160 collects 160 text prompts from four sources: MSCOCO captions [28], DrawBench [44], PartiPrompts [60], and PaintSkill [6]. Each text prompt is paired with five text-to-image models, generating a total of 800 image-text pairs. Furthermore, Davidsonian [5] labels these image-text pairs using 1-5 Likert scale for human evaluation.

Pic-a-pick [24]. We find that the text-to-image evaluation benchmark, Pic-a-pick, contains an excessive amount of NSFW (sexual/violent) content and incorrect labels, likely due to an inadequate automatic filtering procedure. Specifically, after manually reviewing the test set of 500 samples, we find that 10% contain inappropriate content (e.g., "*zentai*" and "*Emma Frost as an alluring college professor wearing a low neckline top*") and approximately 50% had incorrect labels. This may also account for the inferior performance of PickScore. As a result, we manually filter the test set to obtain a clean subset of 100 prompts paired with 200 images for evaluating binary accuracy. We also remove all tied labels due to their subjective nature. We will release this subset of Pick-a-pic for reproducibility.

SeeTrue [59] (DrawBench/EditBench/COCO-T2I). We utilize the binary match-or-not labels collected by SeeTrue [59] for the three benchmarks. These benchmarks consist of individual image-text pairs, where some pairs are correctly paired and others are not. We follow their original evaluation protocols to report the AUROC (Area Under the Receiver Operating Characteristic curve), taking into account all possible classification thresholds.

Winoground [50] and EqBen [53]. In our study, we use the entire Winoground dataset consisting of 400 pairs of image-text pairs. For EqBen, because the official test set includes low-quality images (e.g., very dark or blurry pictures), we analyze the higher-quality EqBen-Mini subset of 280 pairs of image-text pairs, as recommended by their official codebase. These two benchmarks evaluate image-text alignment via matching tasks: each sample becomes 2 image-to-text matching tasks with one image and two candidate captions, and 2 text-to-image matching tasks with one caption and two candidate images. The text (and image) score is awarded 1 point only if *both* matching tasks are correct. The final group score is awarded 1 point only if *all* 4 matching tasks are correct. Importantly, we discover that these benchmarks (especially Winoground) test advanced compositional reasoning skills crucial for understanding real-world prompts, such as counting, comparison, differentiation, and logical reasoning. These advanced compositions operate on basic visual entities, which themselves can be compositions of objects, attributes, and relations.

Table 7. **VQAScore on image-text alignment benchmarks.** We show Group Score for Winoground and EqBen; AUROC for DrawBench, EditBench, and COCO-T2I; pairwise accuracy [11] for TIFA160 and GenAI-Bench; and binary accuracy for Pick-a-Pick, with higher scores indicating better performance for all metrics. VQAScore (based on CLIP-FlanT5) outperforms all prior art across all benchmarks.

Method	Models	Winoground	EqBen	DrawBench	EditBench	COCO-T2I	TIFA160	Pick-a-Pic	GenAI-Bench
Based on vision-language models									
CLIPScore [17]	CLIP-L-14	7.8	25.0	49.1	60.6	63.7	54.1	76.0	51.9
Finetuned on human feedback									
PickScore [24]	CLIP-H-14 (finetuned)	6.8	23.6	72.3	64.3	61.5	59.4	70.0	57.7
ImageReward [58]	BLIPv2 (finetuned)	12.8	26.4	70.4	70.3	77.0	67.3	75.0	57.4
HPSv2 [57]	CLIP-H-14 (finetuned)	4.0	17.0	63.1	64.1	60.3	55.2	69.0	50.1
QG/A methods									
VQ2 [59]	FlanT5, LLaVA-1.5	10.0	20.0	52.8	52.8	47.7	48.7	73.0	53.3
Davidsonian [5]	ChatGPT, LLaVA-1.5	15.5	20.0	78.8	69.0	76.2	54.3	70.0	45.8
VQAScore (ours) using open-source VQA models									
VQAScore	InstructBLIP	28.5	38.6	82.6	75.7	83.0	70.1	83.0	61.9
VQAScore	LLaVA-1.5	29.8	35.0	82.2	70.6	79.4	66.4	76.0	61.6
VQAScore (ours) using our VQA model									
VQAScore	**CLIP-FlanT5**	**46.0**	**47.9**	**85.3**	**77.0**	**85.0**	**71.2**	**84.0**	**63.1**

Table 8. **Failure cases of divide-and-conquer methods (VQ2/Davidsonian).** We show generated question-and-answer pairs of VQ2 and Davidsonian on three GenAI-Bench prompts. These methods often generate irrelevant or erroneous QA pairs (highlighted in red), especially with more compositional texts.

Method	Generated questions	Candidate answers (correct answer choice in bold)
	Text: "a snowy landscape with a cabin, but no smoke from the chimney"	
VQ2	What is the name of the landscape on which it's a cabin?	**a snowy landscape**
	In this landscape what does the fire not go off?	**a cabin**
Davidsonian	Is there a landscape?	**yes**, no
	Is there no smoke from the chimney?	**yes**, no
	Is the cabin in the landscape?	**yes**, no
	Text: "six people wear white shirts and no people wear red shirts"	
VQ2	What does the average American wear?	**white shirts**
	What kind of clothes do not all people wear?	**red shirts**
Davidsonian	Are there people?	**yes**, no
	Are the shirts red?	**yes**, no
	Are the shirts white?	**yes**, no
Text: "in the classroom there are two boys standing together, the boy in the red jumper is taller than the boy in the white t-shirt"		
VQ2	Where do two tall kids stand?	**the classroom**
	Which color of jumper is the tallest?	**the red jumper**
Davidsonian	Is the boy in the red jumper wearing a red jumper?	**yes**, no
	Is the boy in the white t-shirt wearing a white t-shirt?	**yes**, no
	Are the boys standing together?	**yes**, no