Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark

Vitali Petsiuk
Boston University
vpetsiuk@bu.edu

Alexander Siemenn
MIT
asiemenn@mit.edu

Saisamrit Surbehera
Columbia University
ss6365@columbia.edu

Zad Chin
Harvard University
zadchin@college.harvard.edu

Keith Tyser
Boston University
ktyser@bu.edu

Gregory Hunter
Columbia University
geh2129@columbia.edu

Arvind Raghavan
Columbia University
ar4284@columbia.edu

Yann Hicke
Cornell University
ylh8@cornell.edu

Bryan Plummer
Boston University
bplum@bu.edu

Ori Kerret
Ven Commerce
ori@ven.com

Tonio Buonassisi
MIT
buonassi@mit.edu

Kate Saenko
Boston University
saenko@bu.edu

Armando Solar-Lezama
MIT
asolar@csail.mit.edu

Iddo Drori
MIT, Columbia University, Boston University
idrori@csail.mit.edu, idrori@cs.columbia.edu, idrori@bu.edu

Abstract

We provide a new multi-task benchmark for evaluating text-to-image models and perform a human evaluation comparing two of the most common open source (Stable Diffusion) and commercial (DALL-E 2) models. Twenty computer science AI graduate students evaluated the two models, on three tasks, at three difficulty levels, across ten prompts each, providing 3,600 ratings. Text-to-image generation has seen rapid progress to the point that many recent models have demonstrated their ability to create realistic high-resolution images for various prompts. However, current text-to-image methods and the broader body of research in vision-language understanding still struggle with intricate text prompts that contain many objects with multiple attributes and relationships. We introduce a new text-to-image benchmark that contains a suite of fifty tasks and applications that capture a model's ability to handle different features of a text prompt. For example, asking a model to generate a varying number of the same object to measure its ability to count or providing a text prompt with several objects that each have a different attribute to correctly identify its ability to match objects and attributes. Rather than subjectively evaluating text-to-image results on a set of prompts, our new multi-task benchmark consists of challenge tasks at three difficulty levels (easy, medium, and hard) along with human ratings for each generated image.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1 Introduction

Spurred by large-scale pretraining on billions of image-text pairs, vision-language models have seen rapid progress in recent years. Large-scale models like CLIP (1) and Flamingo (2) have reported remarkable performance on dozens of benchmarks using a single model, even when using few or no task-specific training samples. Generating high-resolution images given a text prompt has improved in quality to such an extent with models like Stable Diffusion (3), Imagen (4), and DALL-E 2 (5) that their influence has affected popular culture as illustrated in their use to generate magazine covers[1]. There has been much recent progress in improving text-to-image models, allowing the synthesis of objects within novel contexts (6) such as different backgrounds, illumination, and poses. However, these methods still have challenges generating images in complex scenes or where compositionality is essential. A critical bottleneck in further progress is the lack of rigorous evaluation protocols, as current evaluation methods focus on prompts that do not fully account for the diverse settings these models must support (4).

We propose a new text-to-image generation benchmark covering fifty tasks and applications, each targeting a different capability of text-to-image generation models as shown in Table 2. For example, we may ask a model to produce varying numbers of an object to identify its ability to count or ask a model to generate an image with an object of a specified shape. We divide each task into three difficulty levels: easy, medium, and hard. For example, suppose the task is to synthesize different numbers of objects. In that case, the task may be divided into easy: generating 1-3 objects, medium - generating 4-10 objects, and hard - generating more than ten objects. Next, we provide ten different instances for each task difficulty level. These instances are specific prompts that implement the tasks. We score text-to-image models on each of the thirty instances (ten for each of the three difficulty levels) for each of the fifty tasks and applications. Specifically, we run our benchmark on DALL-E 2 (5) and Stable Diffusion (3).

We can quantify and compare any new text-to-image generation model with our new benchmark. In this work, we perform a human evaluation of three tasks; however, many of the tasks may also be evaluated automatically by a neural network. Table 2 describes which tasks may be evaluated automatically and which require human evaluation. For example, incorporating spatial-aware methods ensures spatial relationships and prompts with compositional elements that are correctly generated. Using OCR mechanisms ensures that quoted text is legible and accurate.

Our key contributions are: (i) developing challenge tasks for state-of-the-art text-to-image generative models, (ii) defining human evaluation procedures and defining which tasks may be automatically evaluated, and (iii) Performing a human evaluation for a subset of tasks with 3,600 human ratings, comparing the performance of two of the most common open source (Stable Diffusion) and commercial (DALL-E 2) models.

1.1 Related Work

Text-to-image models may be roughly split into two types: autoregressive transformer-based models (7) and diffusion-based models (4). Prior state-of-the-art (8; 9; 10; 6) handles specific limitations of text-to-image models such as generating an image within context or modifying object attributes automatically. A comprehensive and quantitative multi-task benchmark for text-to-image synthesis does not exist that covers a diverse set of tasks with varying difficulty levels. Our goal is to develop a benchmark that will become the gold standard in the field for evaluating text-to-image models that will endure the test of time.

Text-to-image models are commonly evaluated by the Inception Score (IS) and the Fréchet Inception Distance (FID). Both of these metrics are based on Inception v3 classifier. These measures, therefore, are designed for the unconditional setting and are primarily trained on single-object images. We have seen several approaches which rectify these shortcomings.

Imagen (4) introduced DrawBench, a benchmark with 11 categories with approximately 200 prompts total. Human raters (25 participants) were asked to choose a better set of generated images from two models regarding image fidelity and image-text alignment. Categories are: colors counting, conflicting, DALL-E 2, description, misspellings, positional, rare words, Reddit, text.

[1] https://www.cosmopolitan.com/lifestyle/a40314356/dall-e-2-artificial-intelligence-cover/

Difficulty	**Easy**		**Medium**		**Hard**		**Average**	
Task / Model	SD	DALL-E 2	SD	DALL-E 2	SD	DALL-E 2	SD	DALL-E 2
Counting	74.8	91.8	52.2	51.4	36.1	54.0	54.4	65.7
Faces	72.5	93.5	74.0	74.3	64.2	77.2	70.2	81.7
Shapes	70.8	67.1	56.8	46.0	45.1	57.3	57.6	56.8

Table 1: Percentages of the best possible scores for human evaluation of Stable Difussion and DALL-E 2 across the three tasks of counting, shapes, and faces. DALL-E 2 outperforms Stable Diffusion on Counting and Faces tasks, Stable Diffusion shows minor advantage on the shapes task. On 6 out of 9 sub-tasks DALL-E 2 produces better images.

DALL-E 2-Eval (11) proposed PaintSkills to test skills of the generative models — specified object generation, counting, color, and spatial relations. It utilizes the Unity engine to test these tasks using predefined sets of objects, a subset of MS-COCO (12) objects, colors, and spatial relations. We propose a more comprehensive benchmark of tasks at a finer level of detail, with three levels of difficulty.

Localized Narratives (13) is a multi-modal image captioning approach that can be adapted to benchmarking images. Text captions are first generated by human annotators whose cursor movement and voice commentary hover their cursor over the image to provide richness and accuracy.

PartiPrompts, a holistic benchmark of 1,600 English prompts (14), compared to Localized Narratives, is better in probing model capabilities on open-domain text-to-image generation. The 1,600 prompts span 12 different categories and 11 challenge aspects. The 12 categories are artifacts, animals, indoor scenes, produce and plants, abstract, arts, food & beverage, vehicles, illustrations, outdoor scenes, people, and world knowledge, while the 11 challenge aspect are basic, fine-grained detail, properties & positioning, linguistic structures, perspective, quantity, writing & symbols, complex, imagination, style & format and simple detail. The image quality and the alignment of the generated image with the input text are evaluated.

2 Methods

We create a comprehensive multi-task text-to-image generation benchmark of fifty diverse tasks and applications, as shown in Table 2. We use the benchmark to compare different models, comparing Stable Diffusion (3) [2] and DALLE-2 (5), and identify their limitations. Our evaluation protocol consists of human ratings between 1 (worst) and 5 (best) of tasks at three difficulty levels.

Examples of cases in which human evaluation is required are: (i) concepts that are difficult to define, such as successfully combining objects that are rarely co-occurring in the real world; (ii) complex tasks such as images that require common sense; and (iii) cases where human expertise is essential such as generating images without racial or gender bias.

We obtained 3,600 human ratings: twenty graduate students, two models, three difficulty levels, and ten prompts each. The images were all generated with identical default model parameters and evaluated on the same scale, providing a rating between 1 (worst) and 5 (best).

3 Results

We survey twenty students that evaluate the performance of Stable Diffusion and DALL-E 2. Each student evaluated the three tasks at three levels of difficulty and ten prompts each, providing a rating between 1 (worst) and 5 (best). We collected a total of 3,600 scores. The results are summarized in Table 1 and a detailed breakdown of the evaluations by prompts is available in Table 3 of the Appendix.

Our human evaluation shows that on the counting task DALL-E 2 (65.7%) performs better than Stable Diffusion (54.4%), on the Faces tasks both models perform very well and DALL-E 2 (81.7%) performs better than Stable Diffusion (70.2%), and on the Shapes task both models perform equally well (56.8% compared to 57.6%).

[2] Publicly available `https://beta.dreamstudio.ai/`

Id	Task	Eval
1	Generating a specified number of objects	ML
2	Generating objects with specified spatial positioning	ML
3	Combining objects very rarely co-occurring in the world	Human
4	Generating images obeying physical rendering rules of shadows, reflections, and acoustics	Human
5	Generating objects with specified colors	ML
6	Generating conflicting interactions between objects	Human
7	Understanding complex and long text prompts describing objects	Human
8	Understanding misspelled prompts	ML
9	Handling absurd requests	Human
10	Understanding rare words	ML
11	Incorporating quoted text with correct spelling	ML
12	Understanding negation and counter-examples	Human
13	Understanding anaphora and phrases that refer to other parts of the prompt	Human
14	Aligning text as specified in the prompt	ML
15	Generating common-sense images	Human
16	Removing objects without needing manual annotation	ML
17	Removing content that is not child-safe	Human
18	Editing the color of objects without marking them manually	ML
19	Replacing objects without marking them manually	ML
21	Objects obeying physics rules	Human
22	Generating images without racial or gender bias	Human
23	Understanding comparative concepts like fewer and more	ML
24	Photo-realistic faces	ML
25	Understanding prompts regarding weather	Human
26	Handling multi-lingual prompts	ML
27	Duplicating objects perfectly	ML
28	Generating multiple camera viewpoints of the same scene	ML
29	Generating realistic faces with a specific emotion	ML
30	Generating well-known faces	ML
31	Generating a thumbnail summary for text and video	Human
32	Changing dimensions of an image without losing information	Human

Id	Application	
33	Graphic designs: generating new designs for websites	
34	e-Commerce: generating personalized ads	
35	Architectural planning: generating new renderings of building and interior designs	
36	Home design: generating creative home design suggestions	
37	Real-estate listings: generating furnished versions of unfurnished apartment and house photos for advertisement	
38	Education: generating personalized digital learning interfaces with customized enhancements	
39	User interface and user experience: generating design templates for mobile and desktop applications	
40	Stop-motion video: generating frames for short animations	
41	Cosmetics: generating realistic images showcasing products	
42	Stock photos: generating large amounts of stock images for general audiences	
43	Product design: quickly prototyping digital and physical products	
44	Illustrations: generating professional artwork for custom purposes	
45	Synthetic data: generating synthetic data for boosting training samples size	
46	Social media: generating memes and shareable content	
47	Image recommendation: generating recommendations based on user preferences	
48	Gaming: using natural language to create complex scenes for video games	
49	Proteomics: designing new proteins visualizing existing structures	
50	Material science: designing new crystals	

Table 2: Multi-task text-to-image benchmark and applications: we propose a series of tasks and applications on which text-to-image models could be evaluated. These tasks can be further divided based on whether the evaluation can currently be automated or requires human assessment.

Performance gracefully degrades as the tasks are more difficult, except for DALL-E 2 which performs slightly better on the hard than on the medium Shapes task.

Our proposed benchmark allows for testing individual competencies and limitations of the different generative models. Understanding the limitations is critical for picking the suitable model for each task and application and advancing the quality of generative models and the alignment of their performance with human goals.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *Image*, vol. 2, p. T2, 2021.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," *arXiv:2204.14198*, 2022.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[6] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," *arXiv preprint arXiv:2208.12242*, 2022.

[7] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, 2022.

[8] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," *arXiv preprint arXiv:2204.02491*, 2022.

[9] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[10] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[11] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers," 2022.

[12] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693. Springer, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48

[13] J. Pont-Tuset, J. R. R. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," *CoRR*, vol. abs/1912.03098, 2019. [Online]. Available: http://arxiv.org/abs/1912.03098

[14] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, “Scaling autoregressive models for content-rich text-to-image generation,” 2022. [Online]. Available: https://arxiv.org/abs/2206.10789

A Appendix

Difficulty	**Easy**		**Medium**		**Hard**	
Prompt / Model	SD	DALL-E 2	SD	DALL-E 2	SD	DALL-E 2
1	96	100	69	32	30	57
2	93	97	62	62	28	52
3	49	97	34	36	22	61
4	87	99	38	46	37	44
5	51	90	77	64	46	57
6	97	50	54	45	35	82
7	92	93	51	79	47	48
8	67	97	38	39	37	45
9	43	98	43	50	30	55
10	73	97	56	61	49	39
1	69	96	79	86	66	73
2	82	93	83	95	70	69
3	76	96	76	75	68	80
4	97	98	70	87	91	91
5	23	92	81	86	58	90
6	75	96	71	77	74	77
7	60	92	91	20	53	95
8	58	84	76	66	41	81
9	96	94	74	77	59	51
10	89	94	39	74	62	65
1	93	92	73	30	36	41
2	94	84	49	36	67	64
3	89	77	32	32	44	62
4	85	49	48	38	44	50
5	49	38	69	42	30	66
6	89	45	45	83	47	71
7	52	80	65	52	34	70
8	37	32	32	42	49	47
9	79	90	84	37	51	37
10	41	84	71	68	49	65
Average:	72.70	84.13	61.00	57.23	48.47	62.83

Table 3: Percentages of the best possible scores for human evaluation of Stable Difussion and DALL-E 2 across the three tasks of counting, shapes, and faces for each of the ten prompts.

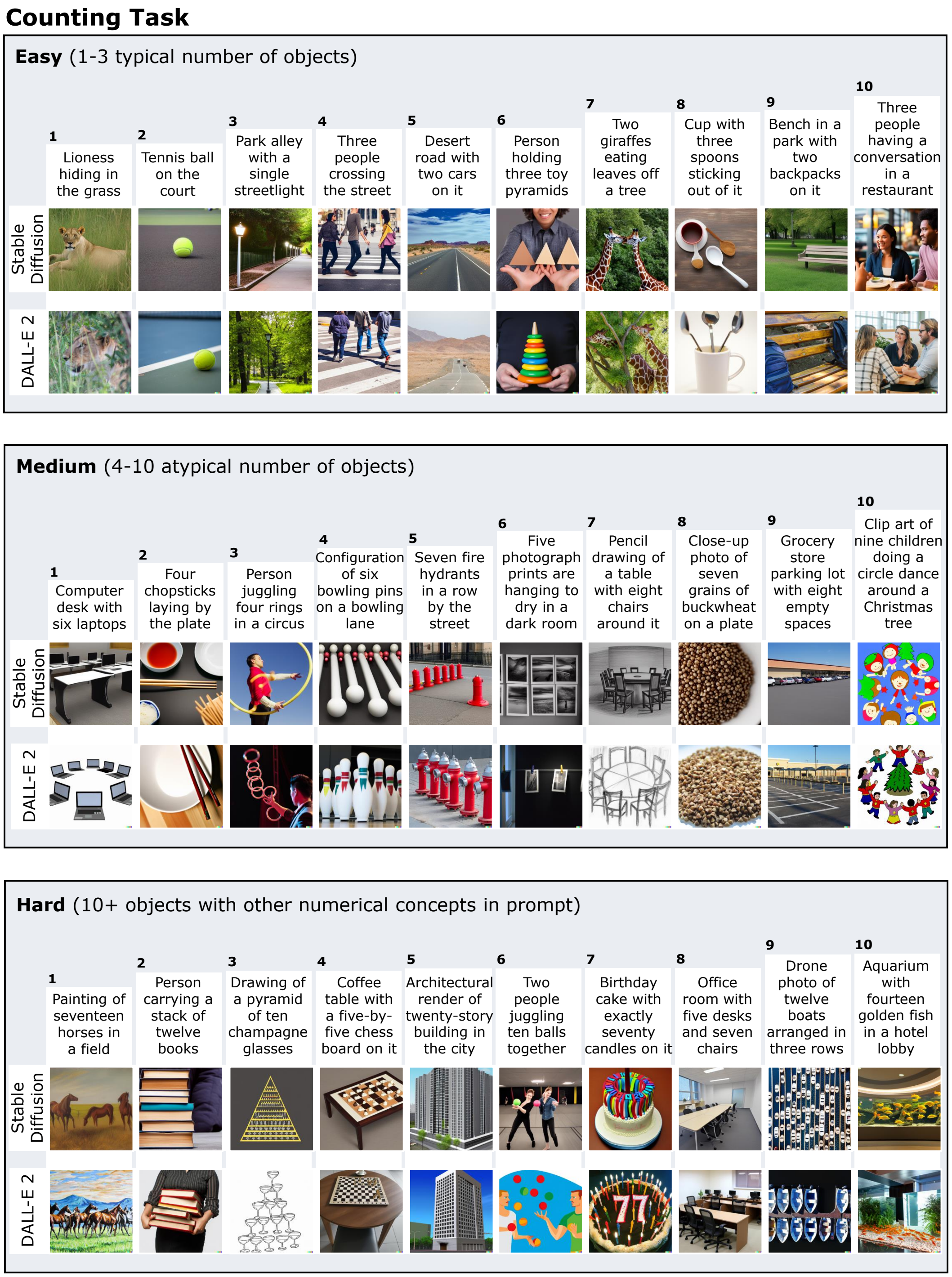

Counting Task
Easy (1-3 typical number of objects)
1 Lioness hiding in the grass
2 Tennis ball on the court
3 Park alley with a single streetlight
4 Three people crossing the street
5 Desert road with two cars on it
6 Person holding three toy pyramids
7 Two giraffes eating leaves off a tree
8 Cup with three spoons sticking out of it
9 Bench in a park with two backpacks on it
10 Three people having a conversation in a restaurant
Stable Diffusion
DALL-E 2
Medium (4-10 atypical number of objects)
1 Computer desk with six laptops
2 Four chopsticks laying by the plate
3 Person juggling four rings in a circus
4 Configuration of six bowling pins on a bowling lane
5 Seven fire hydrants in a row by the street
6 Five photograph prints are hanging to dry in a dark room
7 Pencil drawing of a table with eight chairs around it
8 Close-up photo of seven grains of buckwheat on a plate
9 Grocery store parking lot with eight empty spaces
10 Clip art of nine children doing a circle dance around a Christmas tree
Stable Diffusion
DALL-E 2
Hard (10+ objects with other numerical concepts in prompt)
1 Painting of seventeen horses in a field
2 Person carrying a stack of twelve books
3 Drawing of a pyramid of ten champagne glasses
4 Coffee table with a five-by-five chess board on it
5 Architectural render of twenty-story building in the city
6 Two people juggling ten balls together
7 Birthday cake with exactly seventy candles on it
8 Office room with five desks and seven chairs
9 Drone photo of twelve boats arranged in three rows
10 Aquarium with fourteen golden fish in a hotel lobby
Stable Diffusion
DALL-E 2

Figure 1: Evaluation of image generation on a counting task at various difficulties. Each panel contains tasks at different difficulties where the columns correspond to the text prompt and the rows correspond to the model used to generate the image: (i) Stable Diffusion and (ii) DALLE-2. Images that are evaluated to sufficiently match the prompt have a green border while images that do not sufficiently match the prompt have a red border. A success score for both models is indicated in the upper right corner of each panel. The prompts used to generate the images are classified into three different difficulties: (i) easy difficulty tasks consisting of generating 1–3 objects, *e.g., two cars, three people*; (ii) medium difficulty tasks consisting of generating 4–10 objects, including uncommon combinations of quantities and objects, *e.g., six bowling pins, seven fire hydrants*; (iii) hard difficulty tasks consisting of 10 or more objects with other numerical concepts in the prompt, *e.g., twelve boats arranged in three rows*.

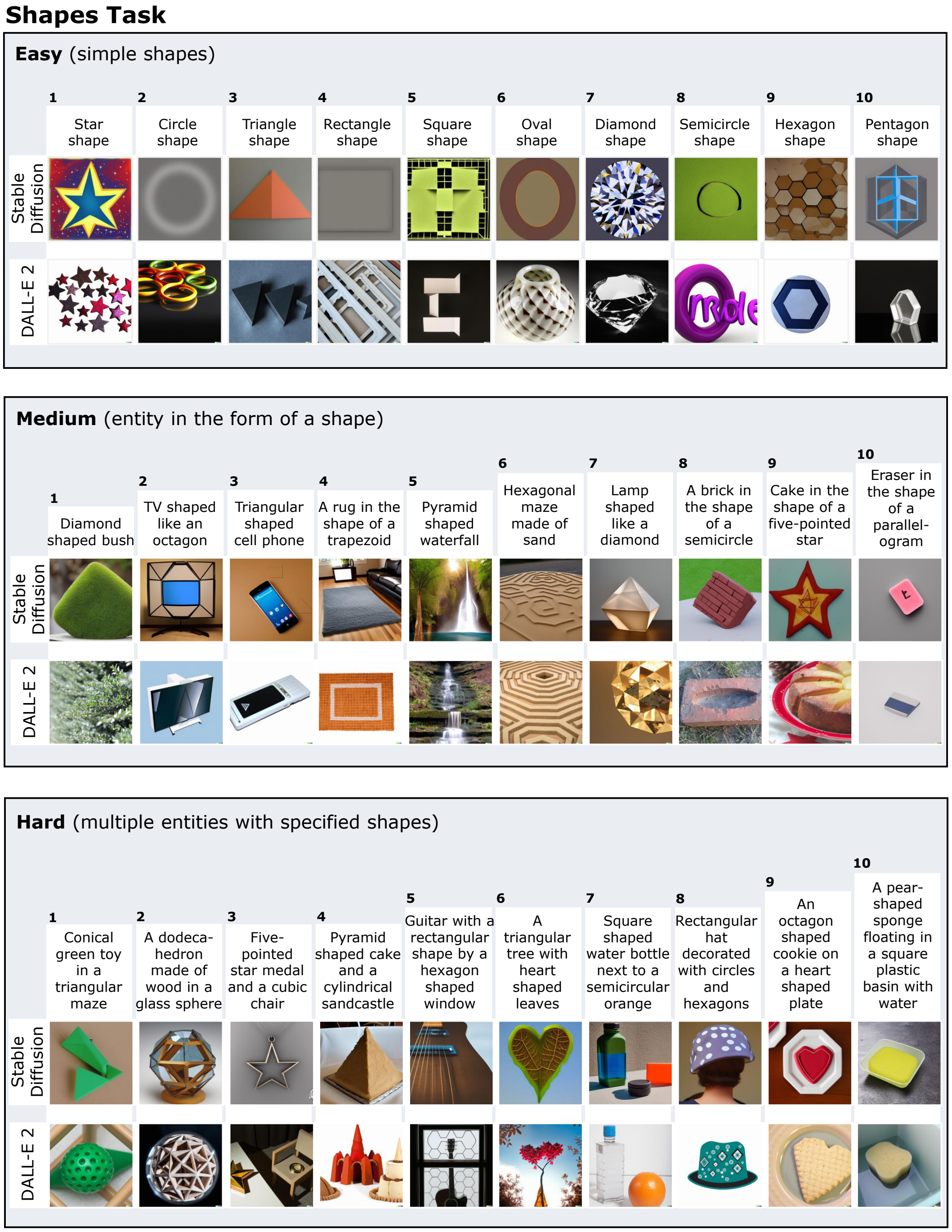
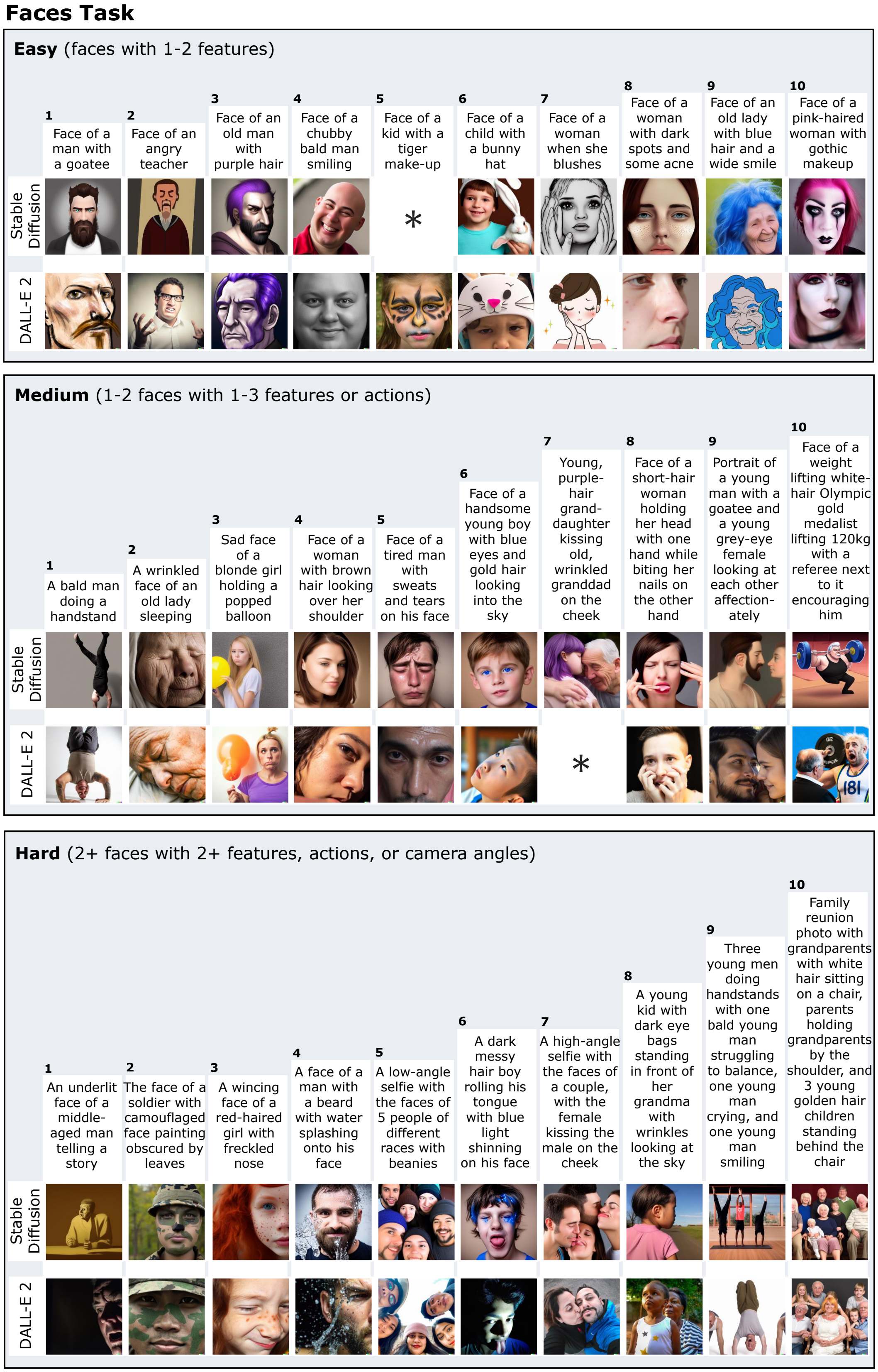

Shapes Task
Easy (simple shapes)
1 Star shape
2 Circle shape
3 Triangle shape
4 Rectangle shape
5 Square shape
6 Oval shape
7 Diamond shape
8 Semicircle shape
9 Hexagon shape
10 Pentagon shape
Stable Diffusion
DALL-E 2
Medium (entity in the form of a shape)
1 Diamond shaped bush
2 TV shaped like an octagon
3 Triangular shaped cell phone
4 A rug in the shape of a trapezoid
5 Pyramid shaped waterfall
6 Hexagonal maze made of sand
7 Lamp shaped like a diamond
8 A brick in the shape of a semicircle
9 Cake in the shape of a five-pointed star
10 Eraser in the shape of a parallel-ogram
Stable Diffusion
DALL-E 2
Hard (multiple entities with specified shapes)
1 Conical green toy in a triangular maze
2 A dodeca-hedron made of wood in a glass sphere
3 Five-pointed star medal and a cubic chair
4 Pyramid shaped cake and a cylindrical sandcastle
5 Guitar with a rectangular shape by a hexagon shaped window
6 A triangular tree with heart shaped leaves
7 Square shaped water bottle next to a semicircular orange
8 Rectangular hat decorated with circles and hexagons
9 An octagon shaped cookie on a heart shaped plate
10 A pear-shaped sponge floating in a square plastic basin with water
Stable Diffusion
DALL-E 2

Figure 2: Evaluation of image generation on a shape task at various difficulties. Each panel contains tasks at different difficulties where the columns correspond to the text prompt and the rows correspond to the model used to generate the image: (i) Stable Diffusion and (ii) DALLE-2. Images that are evaluated to sufficiently match the prompt have a green border while images that do not sufficiently match the prompt have a red border. A success score for both models is indicated in the upper right corner of each panel. The prompts used to generate the images are classified into three different difficulties: (i) easy difficulty tasks consisting of generating simple shapes, *e.g., circle, star*; (ii) medium difficulty tasks consisting of generating entities in the form of a shape, *e.g., hexagonal maze, octagonal TV*; (iii) hard difficulty tasks consisting of multiple entities, each of a specified shape, *e.g., square shaped water bottle next to a semicircular orange*.



Easy (faces with 1-2 features)

1	Face of a man with a goatee
2	Face of an angry teacher
3	Face of an old man with purple hair
4	Face of a chubby bald man smiling
5	Face of a kid with a tiger make-up
6	Face of a child with a bunny hat
7	Face of a woman when she blushes
8	Face of a woman with dark spots and some acne
9	Face of an old lady with blue hair and a wide smile
10	Face of a pink-haired woman with gothic makeup

Medium (1-2 faces with 1-3 features or actions)

1	A bald man doing a handstand
2	A wrinkled face of an old lady sleeping
3	Sad face of a blonde girl holding a popped balloon
4	Face of a woman with brown hair looking over her shoulder
5	Face of a tired man with sweats and tears on his face
6	Face of a handsome young boy with blue eyes and gold hair looking into the sky
7	Young, purple-hair grand-daughter kissing old, wrinkled granddad on the cheek
8	Face of a short-hair woman holding her head with one hand while biting her nails on the other hand
9	Portrait of a young man with a goatee and a young grey-eye female looking at each other affection-ately
10	Face of a weight lifting white-hair Olympic gold medalist lifting 120kg with a referee next to it encouraging him

Hard (2+ faces with 2+ features, actions, or camera angles)

1	An underlit face of a middle-aged man telling a story
2	The face of a soldier with camouflaged face painting obscured by leaves
3	A wincing face of a red-haired girl with freckled nose
4	A face of a man with a beard with water splashing onto his face
5	A low-angle selfie with the faces of 5 people of different races with beanies
6	A dark messy hair boy rolling his tongue with blue light shinning on his face
7	A high-angle selfie with the faces of a couple, with the female kissing the male on the cheek
8	A young kid with dark eye bags standing in front of her grandma with wrinkles looking at the sky
9	Three young men doing handstands with one bald young man struggling to balance, one young man crying, and one young man smiling
10	Family reunion photo with grandparents with white hair sitting on a chair, parents holding grandparents by the shoulder, and 3 young golden hair children standing behind the chair

Figure 3: The prompts used to generate the images are classified into three different difficulties: generating photo-realistic faces given (i) easy: 1 to 2 features for an individual; (ii) medium: different 1 to 3 features for each individual in a group of 1 to 2 people with easy angle and posture; (iii) hard: given more than 2 features for each individual in a group of more than 2 people with challenging angle, posture, lighting or occlusion. Asterisk $(*)$ indicates a failure to generate an image because of model's content filters.