Improving Image Generation with Better Captions

James Betker[*†] jbetker@openai.com
Gabriel Goh[*†] ggoh@openai.com
Li Jing[*†] lijing@openai.com
Tim Brooks[†]

Jianfeng Wang[‡] **Linjie Li**[‡] **Long Ouyang**[†] **Juntang Zhuang**[†] **Joyce Lee**[†] **Yufei Guo**[†]

Wesam Manassra[†] **Prafulla Dhariwal**[†] **Casey Chu**[†] **Yunxin Jiao**[†]

Aditya Ramesh[*†] aramesh@openai.com

Abstract

We show that prompt following abilities of text-to-image models can be substantially improved by training on highly descriptive generated image captions. Existing text-to-image models struggle to follow detailed image descriptions and often ignore words or confuse the meaning of prompts. We hypothesize that this issue stems from noisy and inaccurate image captions in the training dataset. We address this by training a bespoke image captioner and use it to recaption the training dataset. We then train several text-to-image models and find that training on these synthetic captions reliably improves prompt following ability. Finally, we use these findings to build DALL-E 3: a new text-to-image generation system, and benchmark its performance on an evaluation designed to measure prompt following, coherence, and aesthetics, finding that it compares favorably to competitors. We publish samples and code for these evaluations so that future research can continue optimizing this important aspect of text-to-image systems.

1 Introduction

Recent advances in generative modeling have allowed text-to-image generative models to achieve drastic performance improvements. In particular, tackling the problem with sampling-based approaches such as autoregressive generative modeling[27, 2, 1, 20, 30] or using diffusion processes[25, 6, 11, 12, 19, 22] have allowed us to decompose the problem of image generation into small, discrete steps which are more tractable for neural networks to learn.

In parallel, researchers have found ways to build image generators out of stacks of self-attention layers[15, 3, 4]. Decoupling image generation from the implicit spatial biases of convolutions has allowed text-to-image models to reliably improve via the well-studied scaling properties of transformers.

Combined with a sufficiently large dataset, these approaches have enabled the training of large text-to-image models which are capable of generating imagery which is rapidly approaching the quality of photographs and artwork that humans can produce.

[*]Equal contribution
[†]OpenAI
[‡]Microsoft

An outstanding challenge in the field is the controllability of image generation systems, which often overlook the words, word ordering, or meaning in a given caption. We refer to these challenges with the term “prompt following”.

This problem has been pointed out in several works: Rassin et al. (2022) pointed out that DALL-E 2 does not enforce a constraint where each word has a single meaning. Saharia et al. (2022) propose to improve it by conditioning on pre-trained language models, and introduces an evaluation called Drawbench which surfaces common prompt following issues. Yu et al. (2022b) in parallel introduce their own benchmark, Parti Prompts, and show that scaling autoregressive image generators is an alternative way to improve prompt following.

In this work, we propose a new approach to addressing prompt following: caption improvement. We hypothesize that a fundamental issue with existing text-to-image models is the poor quality of the text and image pairing of the datasets they were trained on, an issue that has been pointed out in other works such as Jia et al. (2021). We propose to address this by generating improved captions for the images in our dataset. We do this by first learning a robust image captioner which produces detailed, accurate descriptions of images. We then apply this captioner to our dataset to produce more detailed captions. We finally train text-to-image models on our improved dataset.

Training on synthetic data is not a new concept. For example, Yu et al. (2022b) mention that they apply this technique in training their scaled autoregressive image generators. Our contribution is in building a novel, descriptive image captioning system and measuring the impact of using synthetic captions when training generative models. We also establish a reproducible baseline performance profile for a suite of evals that measure prompt following.

This paper focuses on evaluating the improved prompt following of DALL-E 3 as a result of training on highly descriptive generated captions. It does not cover training or implementation details of the DALL-E 3 model. We provide a high level overview of our strategy for training an image captioner in Section 2, evaluation of text-to-image models trained on original vs. generated captions in Section 3, evaluation of DALL-E 3 in Section 4, and discussion of limitations and risk in Section 5.

In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.

A illustration from a graphic novel. A bustling city street under the shine of a full moon. The sidewalks bustling with pedestrians enjoying the nightlife. At the corner stall, a young woman with fiery red hair, dressed in a signature velvet cloak, is haggling with the grumpy old vendor. the grumpy vendor, a tall, sophisticated man is wearing a sharp suit, sports a noteworthy moustache is animatedly conversing on his steampunk telephone.

Figure 1 – Selected landscape samples from DALL-E 3.

Ancient pages filled with sketches and writings of fantasy beasts, monsters, and plants sprawl across an old, weathered journal. The faded dark green ink tells tales of magical adventures, while the high-resolution drawings detail each creature's intricate characteristics. Sunlight peeks through a nearby window, illuminating the pages and revealing their timeworn charm.

A vibrant 1960s-style poster depicting interplanetary migration, with a retro rocket ship blasting off from earth towards a distant, colorful planet. Bold typography announces "Join the galactic adventure!" with smaller text underneath reading "Explore new worlds, build a brighter future." The background features a swirling galaxy of stars and constellations.

A mischievous ferret with a playful grin squeezes itself into a large glass jar, surrounded by colorful candy. The jar sits on a wooden table in a cozy kitchen, and warm sunlight filters through a nearby window.

A fierce garden gnome warrior, clad in armor crafted from leaves and bark, brandishes a tiny sword and shield. He stands valiantly on a rock amidst a blooming garden, surrounded by colorful flowers and towering plants. A determined expression is painted on his face, ready to defend his garden kingdom.

An icy landscape under a starlit sky, where a magnificent frozen waterfall flows over a cliff. In the center of the scene, a fire burns bright, its flames seemingly frozen in place, casting a shimmering glow on the surrounding ice and snow.

A swirling, multicolored portal emerges from the depths of an ocean of coffee, with waves of the rich liquid gently rippling outward. The portal engulfs a coffee cup, which serves as a gateway to a fantastical dimension. The surrounding digital art landscape reflects the colors of the portal, creating an alluring scene of endless possibilities.

Figure 2 – Selected portrait and square samples from DALL-E 3.

2 Dataset Recaptioning

Our text-to-image models are trained on a dataset composed of a large quantity of pairings (t, i) where i is an image and t is text that describes that image[4]. In large-scale datasets, t is generally derived from human authors who focus on simple descriptions the subject of the image and omit background details or common sense relationships portrayed in image. Important details that are commonly omitted from t might include:

1. The presence of objects like sinks in a kitchen or stop signs along a sidewalk and descriptions of those objects.
2. The position of objects in a scene and the number of those objects.
3. Common sense details like the colors and sizes of objects in a scene.
4. The text that is displayed in an image.

Worse, captions found on the Internet oftentimes simply incorrect; describing tangentially related details of an image. For example, it is common to find advertisements or memes inside of the alt-text commonly used to produce captions for images.

We theorize that all of these shortcomings can be addressed using synthetically generated captions. In subsequent sections, we will discuss the procedure we developed to test out this theory.

2.1 Building an image captioner

An image captioner is very similar to a traditional language model that predicts text. We thus start by providing a brief description of language models. First, a tokenizer is used to break strings of text into discrete tokens. Once decomposed in this way, the text portion of our corpus can be represented as a sequence, $t = [t_1, t_2, \ldots, t_n]$. We can then build a language model over the text by maximizing the following likelihood function:

$$L(t) = \sum_j \log P(t_j | t_{j-k}, \ldots, t_{j-1}; \Theta) \tag{1}$$

Where Θ is the parameters of the captioner that are to be optimized. To turn this language model into a captioner, you need only to condition on the image. The challenge here is that images are composed of many thousands of pixel values. Conditioning on all of this information is exceptionally inefficient with our current neural networks, so we need a compressed representation space. Conveniently, CLIP[17] provides just this.

Thus, given a pre-trained CLIP image embedding function $F(i)$, we augment our language model objective as follows:

$$L(t, i) = \sum_j \log P(t_j | t_{j-k}, \ldots, t_{j-1}; z_j; F(i); \Theta) \tag{2}$$

We follow the methods of Yu et al. (2022a) and jointly pre-train our captioner with a CLIP and a language modeling objective using the above formulation on our dataset of (t, i) text and image pairs. The resulting model is indeed a good captioner, but exhibits the same problems we describe in section 2, such as a reluctance to describe details.

[4]The paired text is generally referred to as the "caption" in this document

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

Figure 3 – Examples of alt-text accompanying selected images scraped from the internet, short synthetic captions (SSC), and descriptive synthetic captions (DSC).

2.1.1 Fine-tuning the captioner

To improve the captions in our image generation dataset, we want to bias our captioner to produce image descriptions which are useful for learning a text-to-image model. In our first attempt, we build a small dataset of captions that describe only the main subject of the image. We then continue to train our captioner on this dataset. The updates to θ induced by this process result in a model which is biased towards describing the main subject of the image. We refer to captions generated by this fine-tune as "short synthetic captions".

We repeat this process a second time, creating a dataset of long, highly-descriptive captions describing the contents of each image in our fine-tuning dataset. These captions describe not only the main subject of the image, but also its surroundings, background, text found in the image, styles, coloration, etc. We again fine-tune our base captioner on this dataset. We refer to captions generated by this captioner as "descriptive synthetic captions".

Figure 3 shows examples of ground-truth, short synthetic, and descriptive synthetic captions.

Once built, we apply our image captioner fine-tunes to every image in our text-to-image dataset, resulting in a set of synthetic captions which we use for subsequent experiments.

3 Evaluating the re-captioned datasets

With our re-captioned datasets in-hand, we set about evaluating the impact of training models on synthetic text. We sought to answer two questions in particular:

1. The performance impact of using each type of synthetic caption.
2. The optimal blending ratio of synthetic to ground-truth captions.

3.1 Blending synthetic and ground-truth captions

Likelihood models like our text-to-image diffusion models have a notorious tendency to overfit to distributional regularities in the dataset. For example, a text-to-image model that is trained on text that always starts with a space character will not work properly if you try to perform inference with prompts that do not also start with that space.

When it comes to training on synthetic captions, we need to consider this issue. Our captioner model could have many modal behaviors that are difficult to detect, but which will become biases of our text-to-image model if it is trained on those captions. Examples of where this might occur is in letter casing, where punctuation appears in the caption (e.g. does it always end with a period?), how long the captions are, or stylistic tendencies such as starting all captions with the words "a" or "an".

The best way to overcome this issue is to regularize our inputs to a distribution of text that is closer to the style and formatting that humans might use. When using ground truth captions, you get this "for free" because these captions are, in fact, drawn from a distribution of human-written text. To introduce some of this regularization into our model training when using synthetic captions, we opted to blend synthetic captions with ground truth captions.

Blending happens at data sampling time, where we randomly select either the ground truth or synthetic caption with a fixed percent chance. We analyze the performance impact of different blending ratios in the next section.

3.2 Evaluation methodology

For evaluation, we trained identical T5-conditioned image diffusion models on the same dataset of images. Details about the models trained are described in A. All models were trained to 500,000 training steps at a batch size of 2048, corresponding to 1B training images total.

Once training was completed, we used captions from an evaluation dataset to generate 50,000 images from each model. We then evaluated these generated images using the CLIP-S evaluation metric outlined in Hessel et al. (2022). We choose CLIP score as a metric as it has a strong correlation with text-image similarity, which is what we are after. As a quick review, this metric is computed as follows:

First, we use the public CLIP ViT-B/32[17] image encoder to produce an image embedding z_i and then we use the text encoder to create a text embedding for the image caption z_t. We finally compute the CLIP score as the cosine similarity C:

$$C(z_i, z_t) = \frac{z_i \cdot z_t}{\|z_i\|\|z_t\|} \tag{3}$$

This metric is then averaged across the distances computed for all 50,000 text/image pairs and rescaled by a factor of 100. We perform this evaluation across multiple model checkpoints during training, at all times performing evaluation with exponentially-weighted averages of the learned weights of the models.

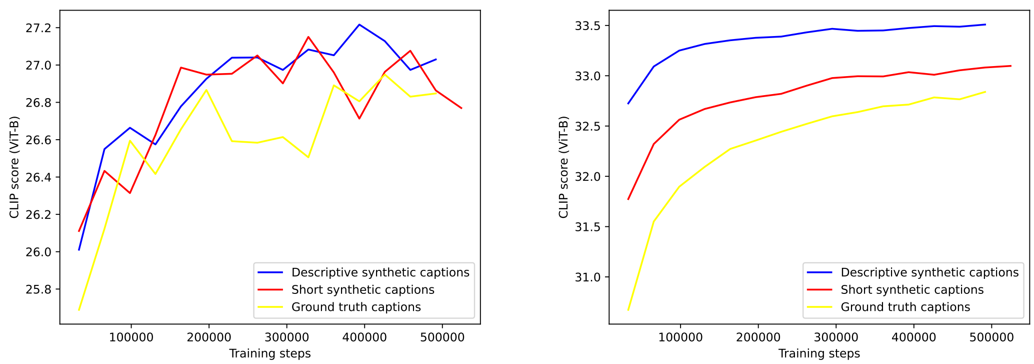


Figure 4 – CLIP scores for text-to-image models trained on different caption types. Left is evaluation results with ground truth captions on our evaluation dataset. Right uses the descriptive synthetic captions from the same dataset.

When computing CLIP scores, the choice of which caption to use when performing the above calculation is important. For our tests, we either use the ground truth caption or we use the descriptive synthetic caption. Which is used is noted in each evaluation.

3.3 Caption type results

We start by analyzing the performance difference between models trained on different types of captions. For this evaluation, we train three models:

1. A text-to-image model trained only on ground truth captions.
2. A text-to-image model trained on 95% short synthetic captions.
3. A text-to-image model trained on 95% descriptive synthetic captions.

We perform this evaluation twice: once with the z_t computed from ground-truth captions and once with z_t computed from descriptive synthetic captions. We do not do it for short synthetic captions as they are very similar to ground-truth in this evaluation.

The results, shown in 4, show that both models trained on synthetic captions achieve slightly better CLIP score performance than the baseline model when evaluated on ground-truth captions, and markedly better performance when evaluated on descriptive synthetic captions. This indicates that there is no downside to using synthetic captions when training text-to-image models.

It is interesting to note that the evaluation curves on synthetic captions have much lower variance. This bolsters our theory that re-captioning can be seen as an averaging operation. Image generation models evaluated on synthetic captions also achieve much higher net CLIP scores across all of the models trained, which supports the notion that synthetic captions have better binding to their corresponding images.

3.4 Caption blending ratios

To evaluate caption blending ratios, we trained four image generation models using our descriptive synthetic captions at different blending ratios. We chose blends of 65%, 80%, 90% and 95% synthetic captions. Midway through the experiment, evaluations showed that the 65% blend was far behind the other blends on all evals and we dropped it.

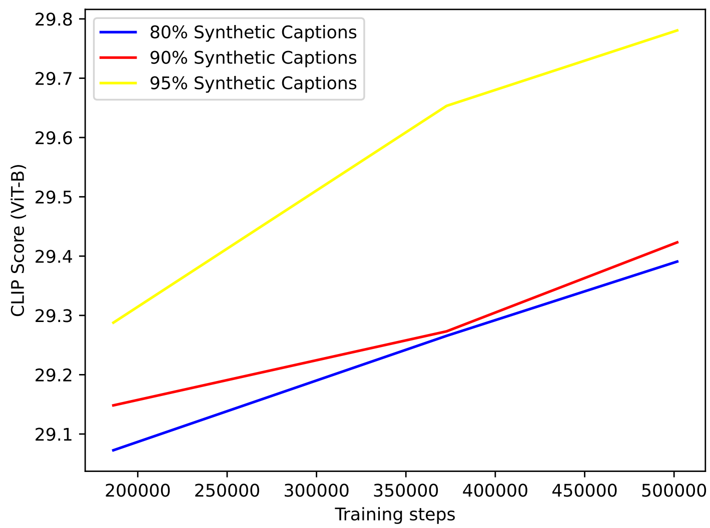


Figure 5 – CLIP scores for text-to-image models trained on various blending ratios of descriptive synthetic captions and ground-truth captions. Evaluation performed using ground truth captions.

The results in Figure 5 show that higher blends of synthetic captions always improved the model's CLIP score.

3.5 Practical usage of highly descriptive captions

The above experiments suggest that we can maximize the performance of our models by training on a very high percentage of synthetic captions. However, doing so causes the models to naturally adapt to the distribution of long, highly-descriptive captions emitted by our captioner.

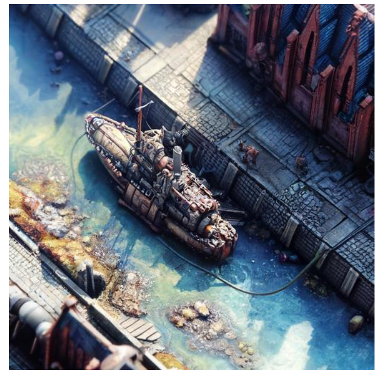
Generative models are known to produce poor results when sampled out of their training distribution. Thus, to extract the maximum potential out of our models, we will need to exclusively sample from them with highly descriptive captions. Fortunately, this is a solvable problem with recent breakthroughs in large language models. Models like GPT-4[14] have become exceptionally good at tasks that require imagination, such as telling stories and writing poems. It stands to reason that they might also be good at coming up with plausible details in an image description.

Indeed, given a prompt such as the one found in Appendix C, we found that GPT-4 will readily "upsample" any caption into a highly descriptive one. To demonstrate how this approach might be useful, we perform this procedure on the captions from the drawbench dataset[24] and visualize the results in Table 7.

As can be seen in Figure 7, utilizing a LLM to "upsample" captions can be used to not only add missing details, but also to disambiguate complex relationships which would be hard for a (relatively) small image generation model to learn. The end result is that the model will often correctly render images that it would have otherwise gotten wrong.

A bird scaring a scarecrow.

Paying for a quarter-sized pizza with a pizza-sized quarter.



A smafml vessef epropoeilled on watvewr by ors, sauls, or han engie.

A large, vibrant bird with an impressive wingspan swoops down from the sky, letting out a piercing call as it approaches a weathered scarecrow in a sunlit field. The scarecrow, dressed in tattered clothing and a straw hat, appears to tremble, almost as if it's coming to life in fear of the approaching bird.

A person is standing at a pizza counter, holding a gigantic quarter the size of a pizza. The cashier, wide-eyed with astonishment, hands over a tiny, quarter-sized pizza in return. The background features various pizza toppings and other customers, all of them equally amazed by the unusual transaction.

A small vessel, propelled on water by oars, sails, or an engine, floats gracefully on a serene lake. the sun casts a warm glow on the water, reflecting the vibrant colors of the sky as birds fly overhead.

Figure 6 – Effect of using "upsampled" drawbench captions to create samples with DALL-E 3. Original drawbench captions on top, upsampled captions on bottom. Images are best of 4 for each caption.

4 DALL-E 3

To test our synthetic captions at scale, we train DALL-E 3, a new state of the art text to image generator. To train this model, we use a mixture of 95% synthetic captions and 5% ground truth captions. The model itself is a scaled-up version of the model we used in the above ablations, with several other improvements.[5]

[5]DALL-E 3 has many improvements over DALL-E 2, many of which are not covered in this document and could not be ablated for time and compute reasons. The evaluation metrics discussed in this document should not be construed as a performance comparison resulting from simply training on synthetic captions.

Metric	DALL-E 3	DALL-E 2 [1]	Stable Diffusion XL[2]
MSCOCO Captions CLIP Score ↑	32.0	31.4	30.5
Drawbench short (GPT-V) [3] ↑	70.4%	49.0%	46.9%
Drawbench long (GPT-V) ↑	81.0%	52.4%	51.1%
T2I-C B-VQA Colors ↑	81.1%	59.2%	61.9%
T2I-C B-VQA Shape ↑	67.5%	54.7%	61.9%
T2I-C B-VQA Texture ↑	80.7%	63.7%	55.2%

Table 1 – Comparison of text-to-image models on various evaluations related to prompt following

[1] Images generated with DALL-E 2 production version live on September 20th, 2023.
[2] Stable Diffusion XL v1.0 with the refiner module active used.
[3] Scores here are the percent of images that have the "correct" caption, as judged by GPT-V.

4.1 Automated Evaluations

We compare DALL-E 3 with DALL-E 2 and Stable Diffusion XL 1.0 with the refiner module[16]. We wish to evaluate DALL-E 3's performance on tasks which are correlated with prompt following. We describe the individual tasks below.

4.1.1 CLIP score

We first compute the CLIP score using the public ViT-B/32 model as described in section 3.2. For this comparison, we use a set of 4,096 captions drawn from the MSCOCO 2014 evaluation dataset[10] to generate our images[6]. In this evaluation, we perform inference on the model with the short, ground-truth captions.

Our model outperforms both DALL-E 2 and Stable Diffusion XL in this evaluation.

4.1.2 Drawbench

We next evaluate on captions from drawbench[24]. For this test, we use an instruction-tuned, vision-aware LLM based on GPT-4 called GPT-V to evaluate the performance of our model versus others. For each prompt in drawbench, we generate four images with each model. We then prompt our vision-aware LLM with the image and the text using the prompt found in Appendix D. This results in a conclusion ("correct"/"incorrect") and an explanation for that conclusion.

Since we previously observed that our model performs better when given extrapolated captions from a language model, we use GPT-4 to "upsample" the drawbench captions using the process described in Section 3.5. We perform the above automated evaluation a second time using these "upsampled" captions when sampling images from all models. We use the original, ground-truth drawbench prompt when asking the vision-aware LLM to judge the outputs.

In all drawbench evaluations, our model beats DALL-E 2 and Stable Diffusion XL. The gap widens significantly when we use the "upsampled" captions.

[6] As with past versions of DALL-E, DALL-E 3 was not specifically trained on the MSCOCO dataset, nor did we perform any optimizations on our model to improve performance on this evaluation. We also did not perform a de-duplication across MSCOCO within our training dataset, it is possible that there is data leakage.

4.1.3 T2I-CompBench

We finally evaluate on a subset the T2I-CompBench evaluation suite developed by Huang et al. (2023). This benchmark measures a model's performance on compositional prompts. We report scores for color binding, shape binding and texture binding. We use the Disentangled BLIP-VQA model to evaluate these results.

DALL-E 3 is state of the art on all benchmarks evaluated.

4.2 Human Evaluations

We submit samples from DALL-E 3 and comparable models for human evaluation. For this evaluation, we present human raters with two side-by-side images that were generated from the same captions. We then ask the rater one of three questions:

1. **Prompt following**: The rater is presented with the full upsampled caption given to the text-to-image model and asked to "choose which image better corresponds to the caption".
2. **Style**: "Imagine you are using a computer tool that produces an image given some text. Choose which image you would prefer to see if you were using this tool."
3. **Coherence**: "Choose which image contains more coherent objects. A "coherent" object is one that could plausibly exist. Look carefully at body parts, faces and pose of humans, placement of objects, and text in the scene to make your judgement. Hint: count instances of incoherence for each image and choose the image with less problems."

For prompt following and style, we assemble a small dataset of 170 captions for this evaluation which is specifically targeted at typical usage of a production text-to-image system. These captions cover a wide array of actual use-cases like generating humans, products and places, concept blending, text rendering and artwork. We call this evaluation set "DALL-E 3 Eval". These captions will be released with our evaluation samples (see Section 4.3). For coherence, we observe that the raters would penalize images depicting imaginary scenes. Therefore, we randomly sample 250 captions from MSCOCO to make sure that the scenes described by the evaluation prompts can plausibly exists. Note that for style and coherence evaluation, we do not show the captions used to generate the images to the raters, to make sure that they would focus on style or coherence, instead of prompt following. For every image pair and question, we poll gather 3 responses from raters, for a total of 2040 ratings per model and question. The human evaluation interface are shown in Section E.

We compare DALL-E 3 versus Stable Diffusion XL with the refiner module and Midjourney v5.2. In Table 2, we report the ELO scores using the same calculation outlined in Nichol et al. (2022).

As the results show, images generated by DALL-E 3 are preferred by human raters a majority of the time over all competitors across all three aspects, especially on prompt following.

Dataset	DALL-E 3	Midjourney 5.2	Stable Diffusion XL	DALL-E 2
DALL-E 3 Eval (prompt following)	**153.3**	-104.8	-189.5	-
DALL-E 3 Eval (style)	**74.0**	30.9	-95.7	-
MSCOCO (coherence)	**71.0**	48.9	-84.2	-
Drawbench	**61.7**	-	-34.0	-79.3

Table 2 – Human evaluation results for DALL-E 3 versus other text-to-image generation models. Scores reported are computed using the ELO algorithm from Nichol et al. (2022)

4.2.1 Drawbench Human Evaluation

In a previous section, we evaluated drawbench using GPT-V. We noticed that for certain types of tests, GPT-V did not exhibit better than random performance in judging prompt following. In particular, this was the case at tasks that involved counting the number of objects in the image. For better coverage of drawbench performance, we submitted images and captions for human evaluation using the procedure described in the previous section. As with our GPT-V drawbench evaluation, we only compare DALL-E 3, Stable Diffusion XL with the refiner module, and DALL-E 2.

4.3 Reproducibility

We have uploaded all of the samples and prompts generated by all models in all of the above comparisons to GitHub.

5 Limitations & Risk

5.1 Spatial awareness

While DALL-E 3 is a significant step forward for prompt following, it still struggles with object placement and spatial awareness. For example, using the words "to the left of", "underneath", "behind", etc are quite unreliable. This is due to the fact that our synthetic captioner also has this weakness: it is unreliable at stating object placement, and this is reflected in our downstream models.

5.2 Text rendering

When building our captioner, we paid special attention to ensuring that it was able to include prominent words found in images in the captions it generated. As a result, DALL-E 3 can generate text when prompted. During testing, we have noticed that this capability is unreliable as words are have missing or extra characters. We suspect this may have to do with the T5 text encoder we used: when the model encounters text in a prompt, it

Photo of a serene park setting. On the left, a golden retriever sits attentively, gazing forward with its tongue out. On the right, a tabby cat lounges lazily, stretching its legs out and looking towards the dog with a curious expression.

Cartoon drawing of an outer space scene. Amidst floating planets and twinkling stars, a whimsical horse with exaggerated features rides an astronaut, who swims through space with a jetpack, looking a tad overwhelmed.

Arum dioscoridis.

Figure 7 – Examples of common failure modes of DALL-E 3

actually sees tokens that represent whole words and must map those to letters in an image. In future work, we would like to explore conditioning on character-level language models to help improve this behavior.

5.3 Specificity

We observed that our synthetic captions are prone to hallucinating important details about an image. For example, given a botanical drawing of a flower, the captioner will often hallucinate a plant genus and species and put it in the caption, even when these details are available in text form in the image. We observed similar behavior when describing pictures of birds: species are hallucinated or not mentioned at all.

This has a downstream impact on our text-to-image models: DALL-E 3 is unreliable at generating imagery for specific terms such as those described above. We believe that further improvements to the captioner should enable further improvements to our text-to-image model.

5.4 Safety and bias mitigations

We performed an in-depth analysis of the safety concerns arising from the deployment of DALL-E 3, including risks posed by model biases. The results of these evaluations can be found in the DALL-E 3 system card[13].

References

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. (2020). Generative pretraining from pixels. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

Chen, X., Mishra, N., Rohaninejad, M., and Abbeel, P. (2017). Pixelsnail: An improved autoregressive generative model.

Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2022). Clipscore: A reference-free evaluation metric for image captioning.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. (2023). T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision.

Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Nichol, A. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

OpenAI (2023a). Dall-e 3 system card.

OpenAI (2023b). Gpt-4 technical report.

Parmar, N., Vaswani, A., Uszkoreit, J., Łukasz Kaiser, Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation.

Rassin, R., Ravfogel, S., and Goldberg, Y. (2022). Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models.

van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., and Kavukcuoglu, K. (2016). Conditional image generation with pixelcnn decoders.

Wu, Y. and He, K. (2018). Group normalization.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022a). Coca: Contrastive captioners are image-text foundation models.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., anda Wei Han, B. H., Parekh, Z., Li, X., Zhang, H., Baldridge, J., and Wu, Y. (2022b). Scaling autoregressive models for content-rich text-to-image generation.

A Image decoder

The image decoder used in our experiments is a text-conditioned U-Net[23] latent diffusion model[22] with three stages.

We use the same VAE[9] developed by Rombach et al. (2022) for our model. This autoencoder performs 8x downsampling. In our synthetic caption evaluation, we train on 256px images, resulting in a model input size of 32x32 latent vectors.

For timestep conditioning, we use a modulated GroupNorm[28], where a learned scale and bias term that is dependent on the timestep signal is applied to the outputs of the GroupNorm layers.

For text-conditioning, we first encode text inputs using a T5 XXL text encoder [18]. The output latents from this process are then cross-attended to by xfnet.

B DALL-E 3 latent decoder

For DALL-E 3, we trained our own diffusion decoder on top of the latent space learned by the VAE trained by Rombach et al. (2022). We found that using a diffusion decoder here provided marked improvements to fine image details, for example text or human faces.

This diffusion decoder is a convolutional U-Net identical to the one described in Ho et al. (2020). Once trained, we used the consistency distillation process described in Song et al. (2023) to bring it down to two denoising steps.

This improved latent decoder is not used for our synthetic caption evaluation.

C Caption "upsampling" prompt for GPT-4

Following is the prompt we give to GPT-4 before feeding it an image caption for "upsampling".

```
You are part of a team of bots that creates images. You work with an assistant bot that will draw anything
you say in square brackets. For example, outputting "a beautiful morning in the woods with the sun peaking
through the trees" will trigger your partner bot to output an image of a forest morning, as described.
You will be prompted by people looking to create detailed, amazing images. The way to accomplish this
is to take their short prompts and make them extremely detailed and descriptive.

There are a few rules to follow:
- You will only ever output a single image description per user request.
- Sometimes the user will request that you modify previous captions. In this case, you should refer to your
  previous conversations with the user and make the modifications requested.
- When modifications are requested, you should not simply make the description longer. You should refactor
  the entire description to integrate the suggestions.
- Other times the user will not want modifications, but instead want a new image. In this case, you should
  ignore your previous conversation with the user.
- Image descriptions must be between 15-80 words. Extra words will be ignored.

    {"role": "user", "content": "Create an imaginative image descriptive caption or modify an earlier
\
caption for the user input: "a man holding a sword\""},
    {"role": "assistant", "content": "a pale figure with long white hair stands in the center of a dark \
    forest, holding a sword high above his head.
the blade glows with a blue light, casting a soft glow on the trees and bushes surrounding him."},
    {"role": "user", "content": "Create an imaginative image descriptive caption or modify an earlier caption \
    for the user input: "make the light red"},
    {"role": "assistant", "content": "a pale figure with long white hair stands in the center of a dark forest, \
    holding a sword high above his head.
the blade glows with a red light, casting a warm glow on the trees and bushes surrounding him."},
    {"role": "user", "content": "Create an imaginative image descriptive caption or modify an earlier caption for the \
    user input: "draw a frog playing dominoes"},
    {"role": "assistant", "content": "a frog sits on a worn table playing a game of dominoes with an elderly raccoon.
the table is covered in a green cloth, and the frog is wearing a jacket and a pair of jeans. The scene is set in a forest,
with a large tree in the background."}
```

D Evaluation of drawbench with vision-enabled GPT-4

Following is the prompt we give to the vision-enabled GPT-4 model to perform our automated drawbench evaluation:

```
You are responsible for judging the faithfulness of images generated by a computer program to the caption
used to generate them. You will be presented with an image and given the caption that was used to produce
the image. The captions you are judging are designed to stress-test image generation programs, and may
include things such as:
1. Scrambled or mis-spelled words (the image generator should an image associated with the probably meaning)
2. Color assignment (the image generator should
apply the correct color to the correct object)
3. Counting (the correct number of objects should be present)
4. Abnormal associations, for example ’elephant under a sea’, where the image should depict what is
    requested.
5. Descriptions of objects, the image generator should draw the most commonly associated object.
6. Rare single words, where the image generator should create an image somewhat associable with the
    specified image.
7. Images with text in them, where the image generator should create an image with the specified text in it.
    You need to make a decision as to whether or not the image is correct, given the caption. You will first
    think out loud about your eventual conclusion, enumerating reasons why the image does or does not
    match the given caption. After thinking out loud, you should output either ’Correct’ or ’Incorrect’ depending
    on whether you think the image is faithful to the caption.
A few rules:
1. Do not nitpick. If the caption requests an object and the object is generally depicted correctly,then
    you should answer ’Correct’.
2. Ignore other objects in the image that are not explicitly mentionedby the caption; it is fine for these to
    be shown.
3. It is also OK if the object being depicted is slightlydeformed, as long as a human would recognize it and
   it does not violate the caption.
4. Your response must always end with either ’incorrect’ or ’correct’
5. ’Incorrect’ should be reserved for instances where a specific aspect of the caption is not followed correctly,
    such as a wrong object, color or count.
6. You must keep your thinking out loud short, less than 50 words.
image(<image_path>)
<prompt>
```

Where `<image_path>` and `<prompt>` are replaced with the image generated by the model and the corresponding prompt used to generate it.

E Human Evaluation Interface

The human evaluation interface for prompt following, coherence and style are shown in Figure 8, 9 and 10.

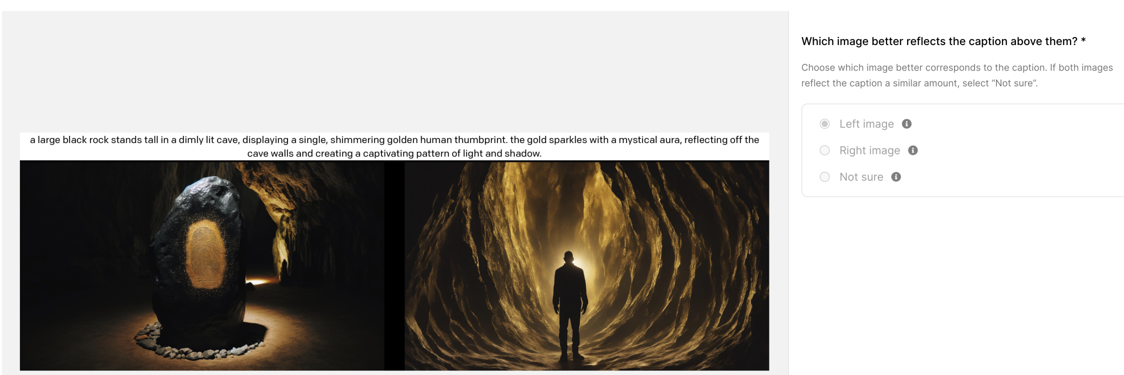


Figure 8 – Human evaluation interface for **prompt following**.

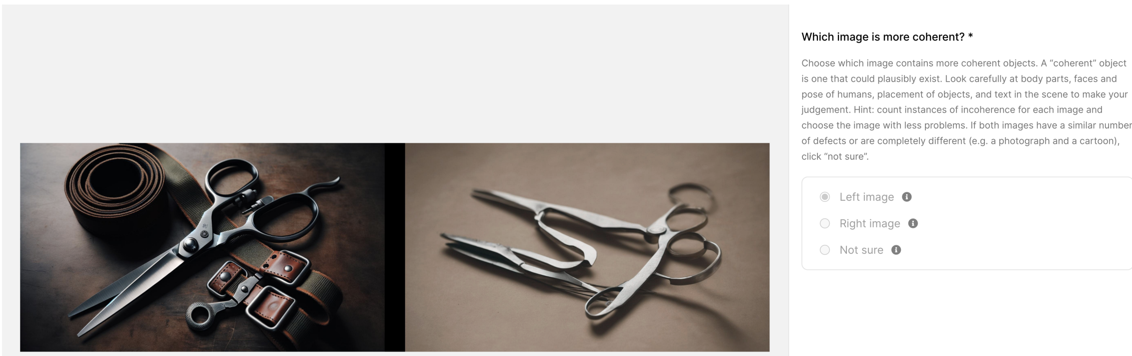


Figure 9 – Human evaluation interface for **coherence**.

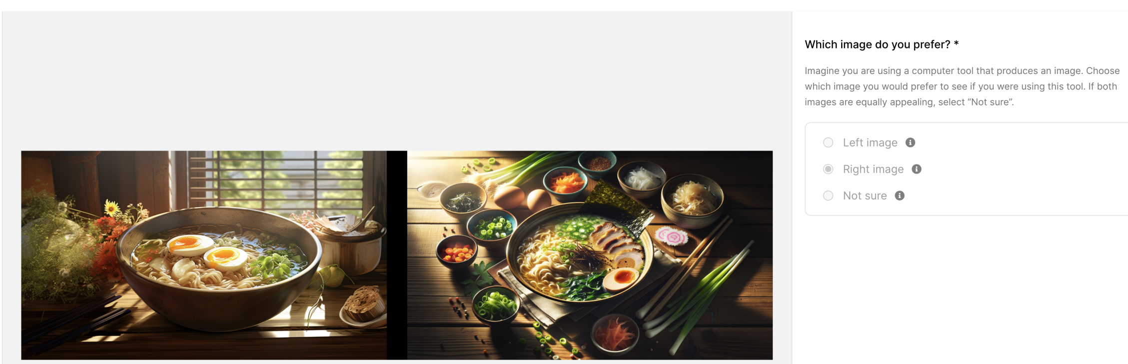


Figure 10 – Human evaluation interface for **style**.