Imagen

unprecedented photorealism × deep level of language understanding

unprecedented photorealism

deep level of language understanding

Google Research, Brain Team

We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g. T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment.

Research Paper DrawBench

A brain riding a rocketship heading towards the moon.

A dragon fruit wearing karate belt in the snow.

Imagen

A small cactus wearing a straw hat and neon sunglasses in the Sahara desert

A photo of a Corgi dog riding a bike in Times Square. It is wearing sunglasses and a beach hat.

Teddy bears swimming at the Olympics 400m Butterfly event.

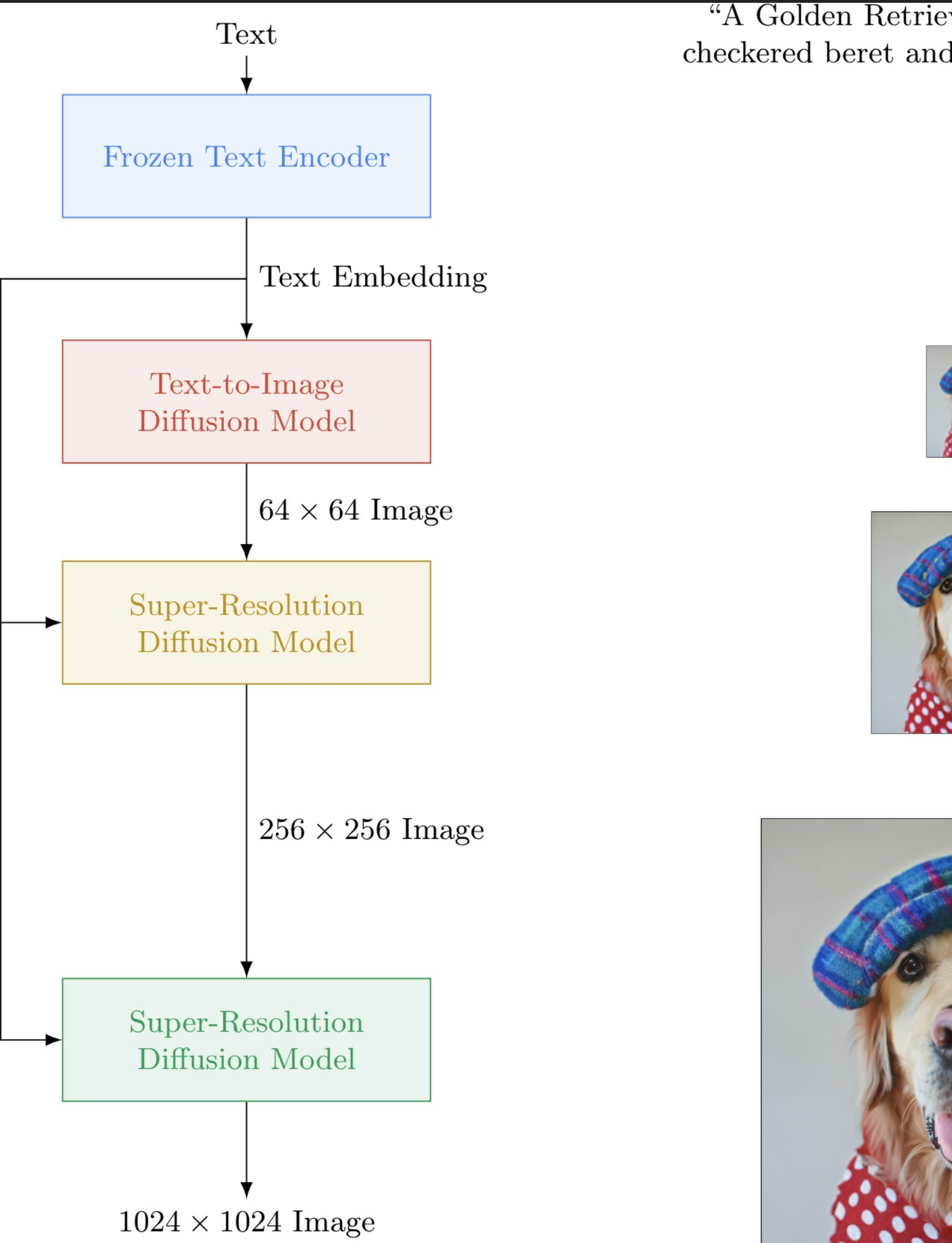
Teddy bears swimming at the Olympics 400m Butterfly event.

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A transparent sculpture of a duck made out of glass. The sculpture is in front of a painting of a landscape.

A single beam of light enters the room from the ceiling. The beam of light is illuminating an easel. On the easel there is a Rembrandt painting of a raccoon.

Imagen is an AI system that creates photorealistic images from input text



Visualization of Imagen. Imagen uses a large frozen T5-XXL encoder to encode the input text into embeddings. A conditional diffusion model maps the text embedding into a 64×64 image. Imagen further utilizes text-conditional super-resolution diffusion models to upsample the image 64×64→256×256 and 256×256→1024×1024.

Large Pretrained Language Model × Cascaded Diffusion Model

deep textual understanding → photorealistic generation

Imagen research highlights

- We show that large pretrained frozen text encoders are very effective for the text-to-image task.
- We show that scaling the pretrained text encoder size is more important than scaling the diffusion model size.
- We introduce a new thresholding diffusion sampler, which enables the use of very large classifier-free guidance weights.
- We introduce a new Efficient U-Net architecture, which is more compute efficient, more memory efficient, and converges faster.
- On COCO, we achieve a new state-of-the-art COCO FID of 7.27; and human raters find Imagen samples to be on-par with reference images in terms of image-text alignment.

Model	COCO FID ↓
Trained on COCO	
AttnGAN (Xu et al., 2017)	35.49
DM-GAN (Zhu et al., 2019)	32.64
DF-GAN (Tao et al., 2020)	21.42
DM-GAN + CL (Ye et al., 2021)	20.79
XMC-GAN (Zhang et al., 2021)	9.33
LAFITE (Zhou et al., 2021)	8.12
Make-A-Scene (Gafni et al., 2022)	7.55
Not trained on COCO	
DALL-E (Ramesh et al., 2021)	17.89
GLIDE (Nichol et al., 2021)	12.24
DALL-E 2 (Ramesh et al., 2022)	10.39
Imagen (Our Work)	7.27

Imagen attains a new state-of-the-art COCO FID.

DrawBench: new comprehensive challenging benchmark

- Side-by-side human evaluation.
- Systematically test for: compositionality, cardinality, spatial relations, long-form text, rare words, and challenging prompts.
- Human raters strongly prefer Imagen over other methods, in both image-text alignment and image fidelity.

State-of-the-art text-to-image

#1 in COCO FID · #1 in DrawBench

Click on a word below and Imagen!

A photo of a An oil painting of a

fuzzy panda British Shorthair cat Persian cat Shiba Inu dog raccoon

wearing a cowboy hat and wearing a sunglasses and

red shirt black leather jacket

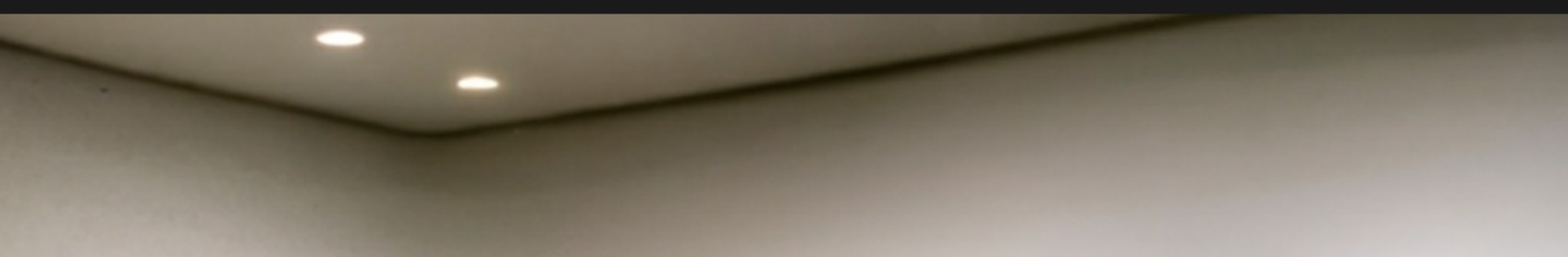
playing a guitar riding a bike skateboarding

in a garden. on a beach. on top of a mountain.

Limitations and Societal Impact

There are several ethical challenges facing text-to-image research broadly. We offer a more detailed exploration of these challenges in our paper and offer a summarized version here. First, downstream applications of text-to-image models are varied and may impact society in complex ways. The potential risks of misuse raise concerns regarding responsible open-sourcing of code and demos. At this time we have decided not to release code or a public demo. In future work we will explore a framework for responsible externalization that balances the value of external auditing with the risks of unrestricted open-access. Second, the data requirements of text-to-image models have led researchers to rely heavily on large, mostly uncurated, web-scraped datasets. While this approach has enabled rapid algorithmic advances in recent years, datasets of this nature often reflect social stereotypes, oppressive viewpoints, and derogatory, or otherwise harmful, associations to marginalized identity groups. While a subset of our training data was filtered to removed noise and undesirable content, such as pornographic imagery and toxic language, we also utilized LAION-400M dataset which is known to contain a wide range of inappropriate content including pornographic imagery, racist slurs, and harmful social stereotypes. Imagen relies on text encoders trained on uncurated web-scale data, and thus inherits the social biases and limitations of large language models. As such, there is a risk that Imagen has encoded harmful stereotypes and representations, which guides our decision to not release Imagen for public use without further safeguards in place.

Finally, while there has been extensive work auditing image-to-text and image labeling models for forms of social bias, there has been comparatively less work on social bias evaluation methods for text-to-image models. A conceptual vocabulary around potential harms of text-to-image models and established metrics of evaluation are an essential component of establishing responsible model release practices. While we leave an in-depth empirical analysis of social and cultural biases to future work, our small scale internal assessments reveal several limitations that guide our decision not to release our model at this time. Imagen, may run into danger of dropping modes of the data distribution, which may further compound the social consequence of dataset bias. Imagen exhibits serious limitations when generating images depicting people. Our human evaluations found Imagen obtains significantly higher preference rates when evaluated on images that do not portray people, indicating a degradation in image fidelity. Preliminary assessment also suggests Imagen encodes several social biases and stereotypes, including an overall bias towards generating images of people with lighter skin tones and a tendency for images portraying different professions to align with Western gender stereotypes. Finally, even when we focus generations away from people, our preliminary analysis indicates Imagen encodes a range of social and cultural biases when generating images of activities, events, and objects. We aim to make progress on several of these open challenges and limitations in future work.



An art gallery displaying modern paintings. The art gallery is flooded. Robots are floating around the art gallery using paddle boards.

A majestic oil painting of a raccoon Queen wearing red French royal robe. The painting is hanging on a wall decorated with wallpaper. A

A blue jay standing on a large basket of rainbow macarons.

A blue jay standing on a large basket of rainbow macarons.

A giant cobra snake on a farm. The snake is made out of corn.

Imagen

imagine · illustrate · inspire

Authors

Chitwan Saharia[*], William Chan[*], Saurabh Saxena[†], Lala Li[†], Jay Whang[†], Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho[†], David Fleet[†], Mohammad Norouzi[*]

[*]Equal contribution. [†]Core contribution.

Special Thanks

We give thanks to Ben Poole for reviewing our manuscript, early discussions, and providing many helpful comments and suggestions throughout the project. Special thanks to Kathy Meier-Hellstern, Austin Tarango, and Sarah Laszlo for helping us incorporate important responsible AI practices around this project. We appreciate valuable feedback and support from Elizabeth Adkison, Zoubin Ghahramani, Jeff Dean, Yonghui Wu, and Eli Collins. We are grateful to Tom Small for designing the Imagen watermark. We thank Jason Baldridge, Han Zhang, and Kevin Murphy for initial discussions and feedback. We acknowledge hard work and support from Fred Alcober, Hibaq Ali, Marian Croak, Aaron Donsbach, Tulsee Doshi, Toju Duke, Douglas Eck, Jason Freidenfelds, Brian Gabriel, Molly FitzMorris, David Ha, Philip Parham, Laura Pearce, Evan Rapoport, Lauren Skelly, Johnny Soraker, Negar Rostamzadeh, Vijay Vasudevan, Tris Warkentin, Jeremy Weinstein, and Hugh Williams for giving us advice along the project and assisting us with the publication process. We thank Victor Gomes and Erica Moreira for their consistent and critical help with TPU resource allocation. We also give thanks to Shekoofeh Azizi, Harris Chan, Chris A. Lee, and Nick Ma for volunteering a considerable amount of their time for testing out DrawBench. We thank Aditya Ramesh, Prafulla Dhariwal, and Alex Nichol for allowing us to use DALL-E 2 samples and providing us with GLIDE samples. We are thankful to Matthew Johnson and Roy Frostig for starting the JAX project and to the whole JAX team for building such a fantastic system for high-performance machine learning research. Special thanks to Durk Kingma, Jascha Sohl-Dickstein, Lucas Theis and the Toronto Brain team for helpful discussions and spending time Imagening!