

# Рецензия

## Авторы

Статью выпустили Google Brain в 2022 году. У нее довольно много авторов, поэтому давайте выделим самых интересных.

Jonathan Ho - его имя мы уже слышали на HICe, это один из авторов DDPM.

Тим Салиманс, до Google Brain какое-то время проработал в OpenAI, и является одним из авторов первой версии GPT. А еще в свое время он применял RL к доте 2.

Вместе с Тимом Джонатан написал статью про classifier free guidance, эта идея многократно переиспользовалась, в том числе и в Imagen.

Вместе с группой людей Джонатан и Тим также работали над статьей про каскадные диффузионки, что так же лежит в основе Imagen. Так получилось, что все авторы этой статьи вошли в авторский состав Imagen. Вот самые интересные личности среди них:

Вильям Чан - автор Listen Attend and Spell, SpecAugment и WaveGrad.

И Мухамад Норузи, который так же работал над WaveGrad и еще является автором SimCLR.

## Предпосылки

Мы уже знаем, что в 2020 году вышла статья DDPM, которая можно сказать положила начало всем последующим статьям про диффузионки в генерации картинок.

Но text-to-image диффузионки захватили не сразу. Еще в начале 2021 году вышла первая DALL-E, в основе которой лежат трансформер и вариационный автоэнкодер, и VQ-GAN.

Но уже в конце 2021 появились первые использовать диффузионки: GLIDE от OpenAI и Latent Diffusion Models.

Авторы GLIDE стали первыми, кто попробовал применить guidance для диффузии, используя для этого текст.

В LDM представили интересную идею учить диффузию на скрытых представлениях из автоэнкодера. Авторы использовали модификацию UNet с кросс-вниманием внутри блоков, и эту идею переиспользовали авторы Imagen.

## Imagen VS DALL-E 2

А затем весной 2022 года практически одновременно вышли две конкурирующие разработки: DALL-E 2 от OpenAI и наш Imagen.

Несмотря на свое название, DALL-E 2 по сути представляет собой модификацию модели GLIDE, нежели прямое продолжение своей первой версии.

То есть это диффузия, которая чуть хитрее использует CLIP guidance по сравнению с GLIDE.

Есть пара моментов, в которых DALL-E 2 уступает Imagen.

Во-первых, в DALL-E 2 текстовый энкодер - из отдельно обученного CLIPa. Обучается он с нуля на тех же парах текст-картинка, что и диффузия. Использование текстового энкодера, предобученного на куче текста, как в Imagen, приводит к лучшему пониманию текста и к более точным генерациям.

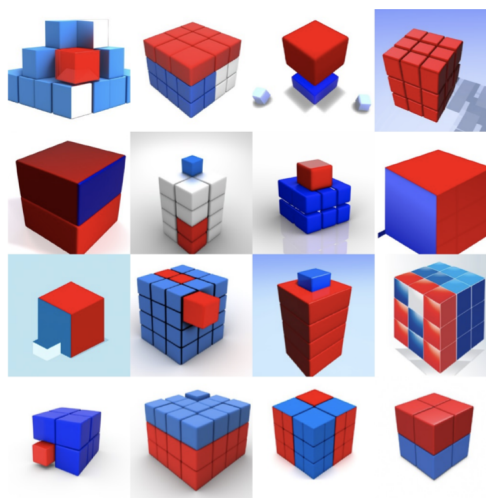
Кроме того, DALL-E 2 не умеет генерировать текст на картинках и если ее попросить нарисовать несколько объектов разных цветов, или как-то конкретно расположенные относительно друг-друга, она запутается.

Существует мнение, что во всем виноват CLIP. Его текстовый энкодер не способен закодировать информацию о соотношениях объект-цвет и их взаимном расположении. А также он не кодирует написание полученного текста. Отсюда все ошибки.

Примеры ошибок DALL-E 2:



Prompt: A sign that says deep learning

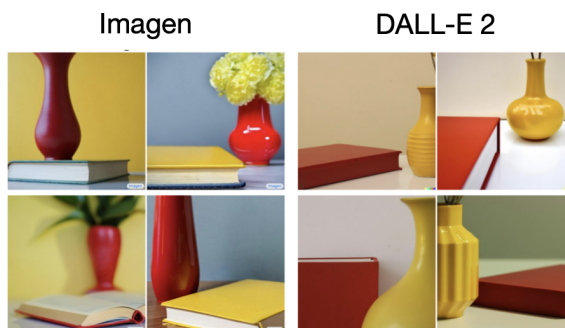


Prompt: A red cube on top of a blue cube

И как их победил Imagen:



Prompt: A storefront with Text to Image written on it



Prompt: A yellow book and a red vase

## Продолжение

Теперь про то, что было дальше.

Есть практически прямое продолжение Imagen - Imagen Video. Те же люди взяли те же идеи и видео-диффузию и получили генерацию видео по промптам.

DeepFusion - тоже разработка гугла, способная генерировать 3д модели по промптам. Это по сути Imagen, над которым надстроили что-то похожее на NeRF.

И еще, Stable Diffusion - в основу тут легли Latent Diffusion Models. Но используется идея с предобученным замороженным текстовым энкодером, как в Imagen.

## Популярность










Может возникнуть вопрос, раз Imagen так крут, то почему все обсуждает DALL-E 2 и Stable Diffusion, а его нет?

Тут причина довольно в стиле гугла - они решили не делать Imagen публичным. В статье целый раздел посвящен описанию различных опасений авторов, связанных с недостаточно приемлемыми данными. Еще, например, авторы исследовали свою модель и поняли, что она более успешно генерирует людей со светлой кожей и иллюстрации стереотипных представлений о людях.

Довольно интересный раздел, и здорово что они сами признают, что учились на непонятных данных с кучей картинок из раздела Неприемлемый контент.

Но жаль, что из-за этого они проиграли гонку на узнаваемость.

## Источники

-  [What are Diffusion Models?](#)
-  [Diffusion models explained. How does OpenAI's GLIDE work?](#)
-  [12 Must read Text to Image AI Research Papers with their code implementation](#)
-  [Google's Imagen AI: Outrageously Good!](#)
-  [Text-to-Image is All the Rage. So Why Aren't We Talking About Imagen?](#)
-  [Is it better than DALL-E 2? | How does Imagen Actually Work?](#)
-  [OpenAI and the road to text-guided image generation: DALL·E, CLIP, GLIDE, DALL·E 2 \(unCLIP\)](#)
-  [Google's Imagen vs OpenAI's DALL-E-2](#)
-  [Guidance: a cheat code for diffusion models](#)