Generation of Any Visual Data with Text2Image Diffusion Models: A Mirage or the Ultimate Reality

June, 2023

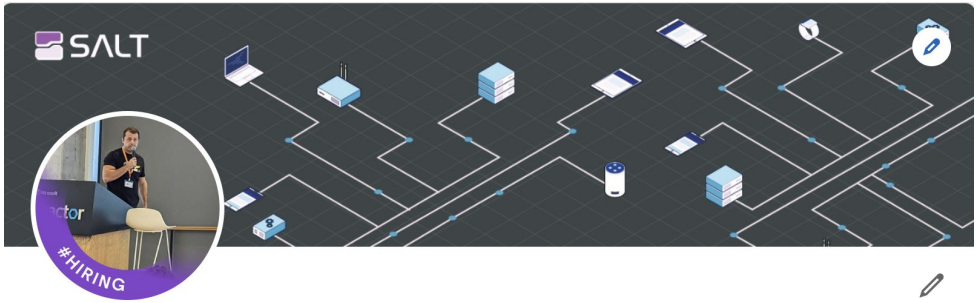
Michael(Mike) Erlihson, PhD

Outline:

- Denoising Diffusion Probabilistic Models(DDPMs or DMs): An Essence
- Popular Text2Image DMs(Imagen, DallE-2, Stable Diffusion, MidJourney)
- 4 Broad Approaches to Leverage Text2Image DMs to generate visual data
- Leveraging Text2Image DMs for Text2Video
- Text2Image Models for Image and Video Editing (Personalization)
- Building and Editing 3D Models and Point Clouds with Text2Image Model
- Controlling your Text2Image Model (edges, segmentation, depth map,...)

Some ML-related info about Mike 😎

- Principal DS at Salt Security
  - Fighting API attacks with some advanced ML
- #deepnightlearners Founder
  - Deep Learning Paper Reviews in Hebrew and in English(> 70 published reviews)
- Machine and Deep Learning in Hebrew Book: Coauthor



Denoising Diffusion Probabilistic Models:

A New Generative Computer Vision Boss

How did it started?

Computer Science > Machine Learning

[Submitted on 11 May 2021 (v1), last revised 1 Jun 2021 (this version, v4)]

Diffusion Models Beat GANs on Image Synthesis

CODE

Prafulla Dhariwal, Alex Nichol

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at this https URL

Comments: Added compute requirements, ImageNet 256×256 upsampling FID and samples, DDIM guided sampler, fixed typos
Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV); Machine Learning (stat.ML)
Cite as: arXiv:2105.05233 **[cs.LG]**
(or arXiv:2105.05233v4 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2105.05233

Submission history

From: Prafulla Dhariwal [view email]
[v1] Tue, 11 May 2021 17:50:24 UTC (30,977 KB)
[v2] Wed, 12 May 2021 17:57:59 UTC (30,978 KB)
[v3] Thu, 13 May 2021 17:57:08 UTC (30,980 KB)
[v4] Tue, 1 Jun 2021 17:49:49 UTC (44,152 KB)

This paper has ignited an intense race in the CV community

to develop DM-powered models across all domains and tasks

Video Generation, 3D Models/Point Clouds, Image/Video Editing, Object Detection, Segmentation…

Super cool DM-featured Text-to-Image (T2I) generative models started popping up…

Dall-E2, Imagen, Stable Diffusion, MidJourney….

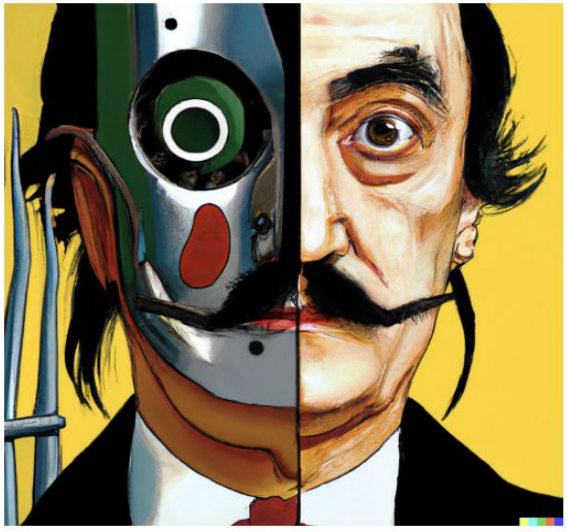
The results were just stunning!

Dalle-2, OpenAI: T2I Diffusion Model

A dolphin in an astronaut suit on saturn, artstation

A panda mad scientist mixing sparkling chemicals, artstation

A vibrant portrait painting of Salvador Dalí with a robotic half face

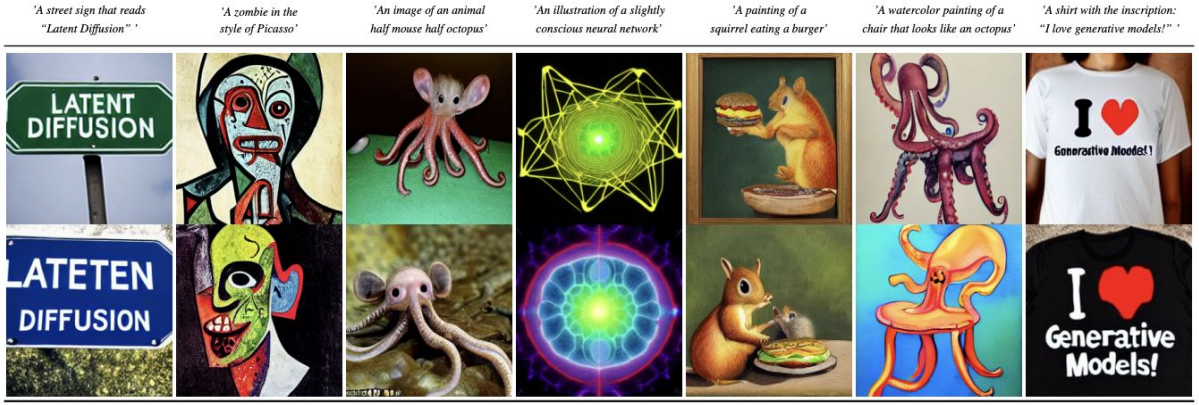


Imagen, Google Research: Longer Texts

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

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

Stable Diffusion: Stability AI



futuristic house in a suburban neighborhood nighttime serious moody

3D model painting of a human cyborg in a city salvador dali neon 8K highly detaled

3D model painting of a human cyborg in a city picasso neon 8K highly detaled

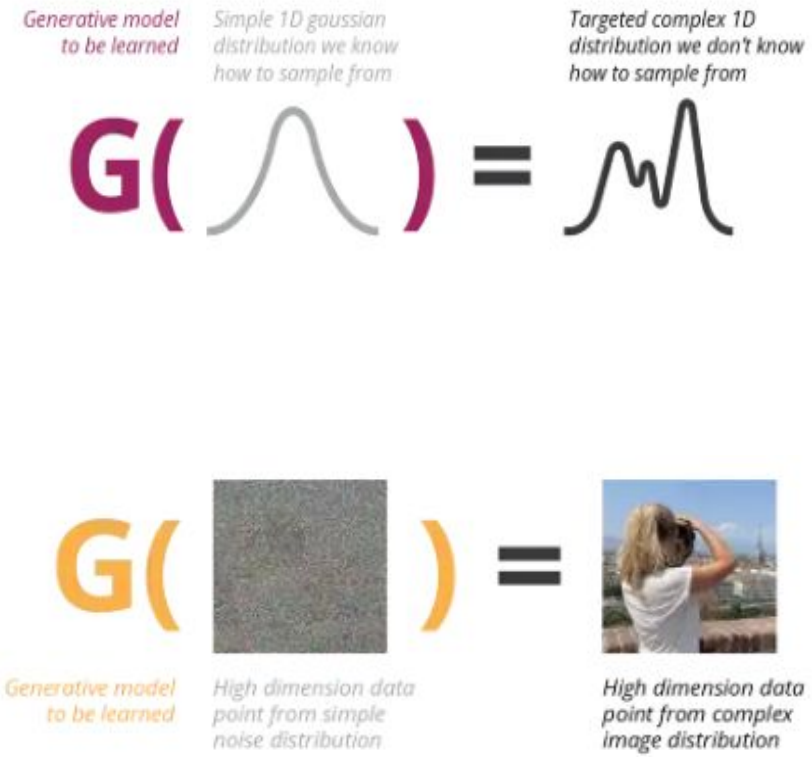
Can we **Harness** these generative Text2image models

to **Generate** other types of visual data?

YES!!

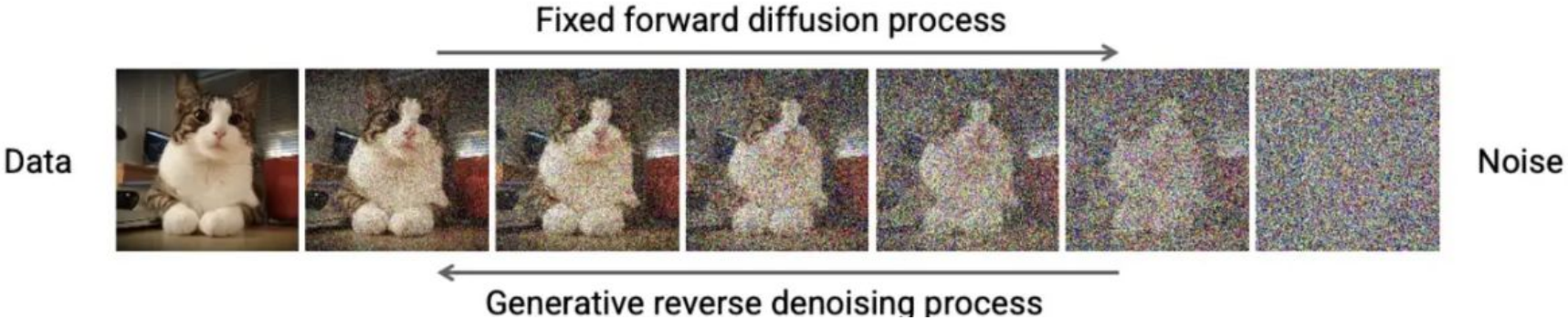
3D models, video generation, image/video editing/personalization, or even Semantic Segmentation

Modern Generative Models: An Essence



Diffusion Model: An Essence

- Destroy data by gradually adding of Gaussian noise (forward process)
- Learn to recover the data by reversing the forward process(backward process)

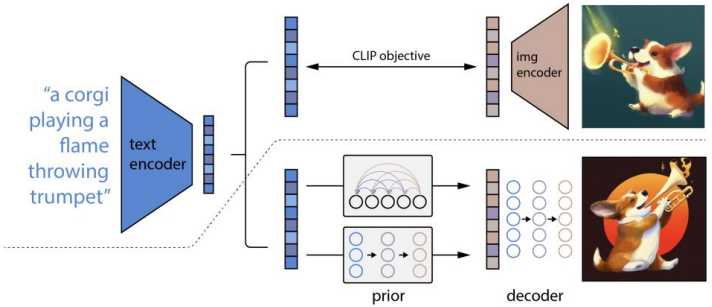


Diffusion Model: An Essence

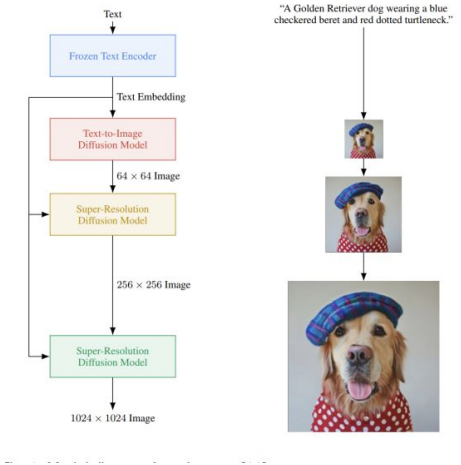
Once DM has learned how to model the reverse process

We can feed **random gaussian noise** into it to generate new pieces of data

Text-to-Image DMs: Model Architectures

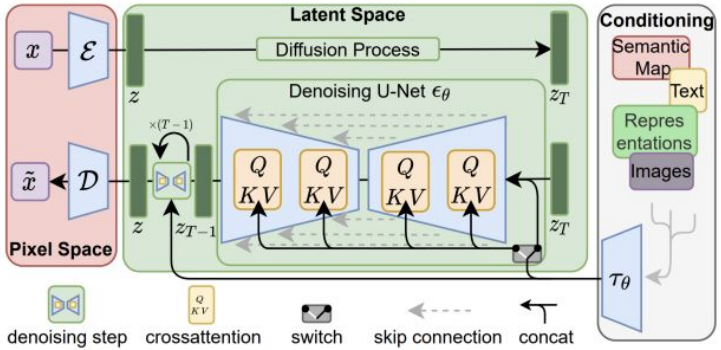


DallE-2



Imagen

Stable Diffusion



Text2Image DMs: Align Text and Image Representations

Architecture Type 1: Imagen, DallE-2

- Use LLM to encode a textual description
- Map Text Embedding into an Image Embedding (DM) - optional
- Use text or/with image embeddings as an input to a **pixel** Text2Image DM

Text2Image DMs: Align Text and Image Representations

Architecture Type 2, Stable Diffusion

- Train image encoder/decoder (VQ-GAN) and a text encoder
- DM is trained **in the embedding space** with text embedding as an input
- The DM output is “decoded” by the decoder to generate an image

Exploit Pretrained Text2Image DMs to Generate Visual Data

Main Approaches:

- Modify output format of DMs to create different types of data
  - Video creation for a given text description

- Fine-tuning DMs or “choosing” tokens for your object: personalization
  - Generate images of a specific dog/car/watch….

Exploit Pretrained Text2Image DMs to Generate Visual Data

Main Approaches:

- Use pretrained DMs to control (propagate) quality/alignment of visual models “projections” with the textual description
  - Views for 3D Models/Point Cloud

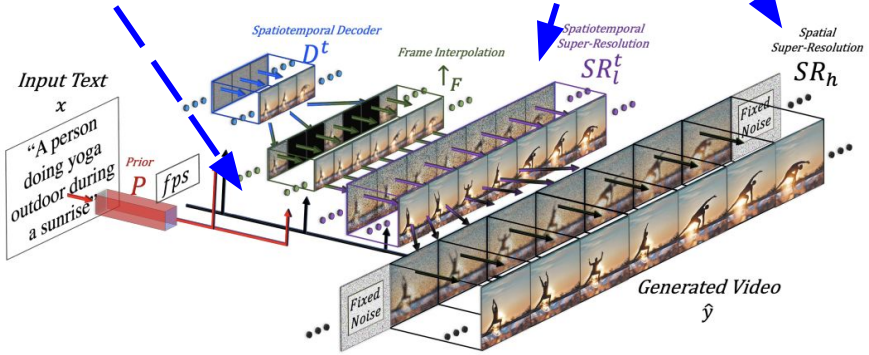
- Adapt pretrained DMs to receive input of different types (in addition to text)
  - Canny edges, Hough lines, user scribbles, human key points,... (ControlNet)

Modify output format of DMs: Make-A-Video

Text2Video: DM generates multiple images

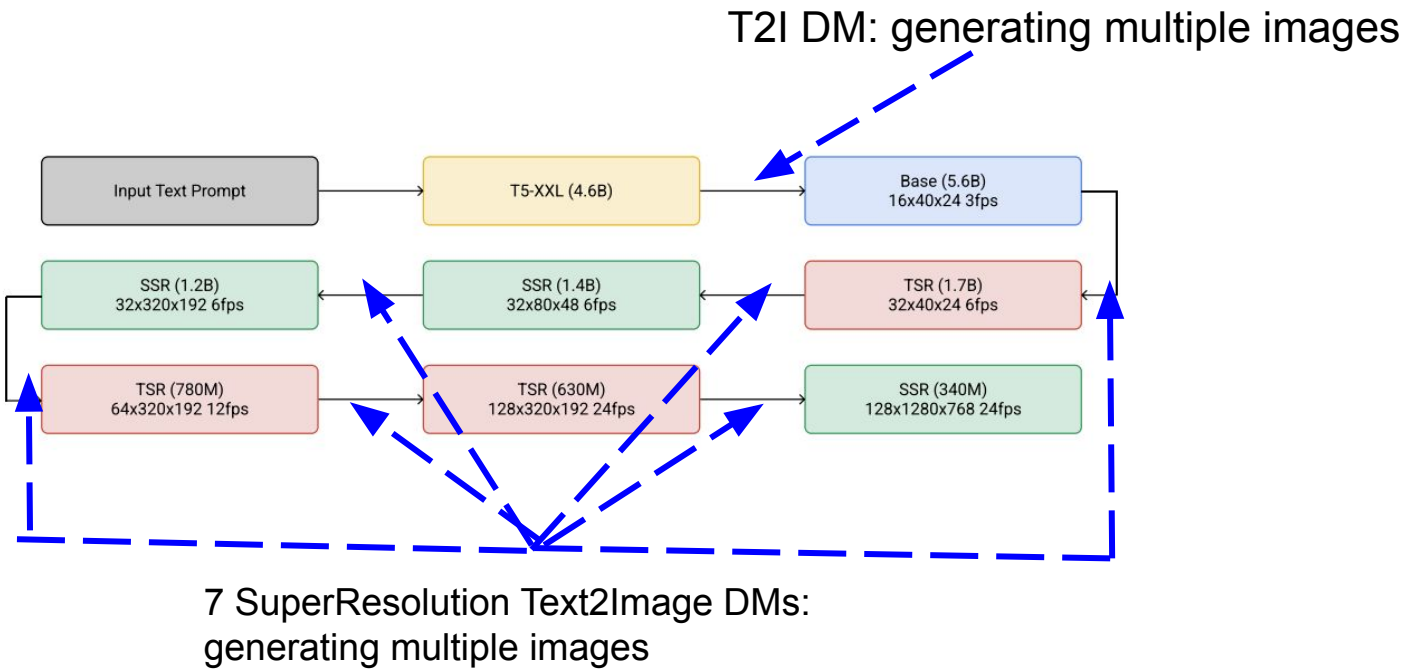
SuperResolution T2I DMs:
generating multiple images/frames

Text2Image DM:
generating multiple images



Modify output format of DMs: Imagen Video

Text2Video: DM generates multiple images



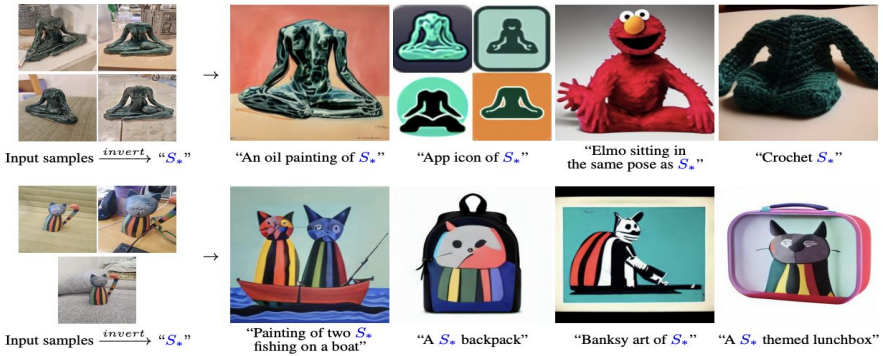
Personalization: choose tokens for your object

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Rinon Gal[1,2*] Yuval Alaluf[1] Yuval Atzmon[2] Or Patashnik[1]

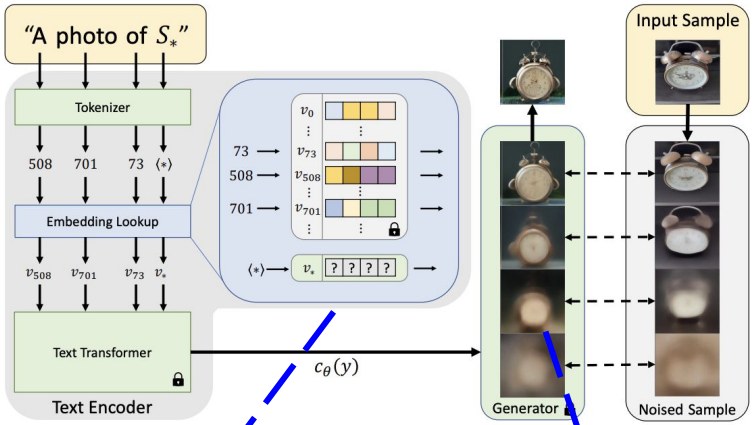
Amit H. Bermano[1] Gal Chechik[2] Daniel Cohen-Or[1]

[1]Tel-Aviv University [2]NVIDIA



Text2Image DM is fine-tuned per object

Personalization: “choose” an object name



“Choose” tokens for your object → optimize over the token space

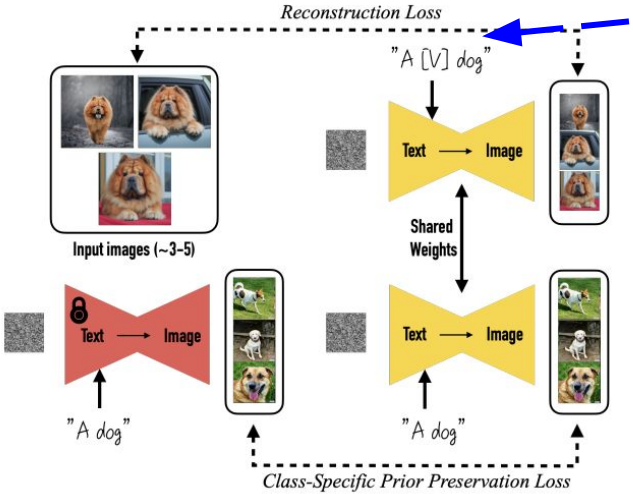
Frozen Text2Image DM

Personalization: fine-tune Text2Image DM for your object

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz*,[1,2] Yuanzhen Li[1] Varun Jampani[1]
Yael Pritch[1] Michael Rubinstein[1] Kfir Aberman[1]
[1] Google Research [2] Boston University

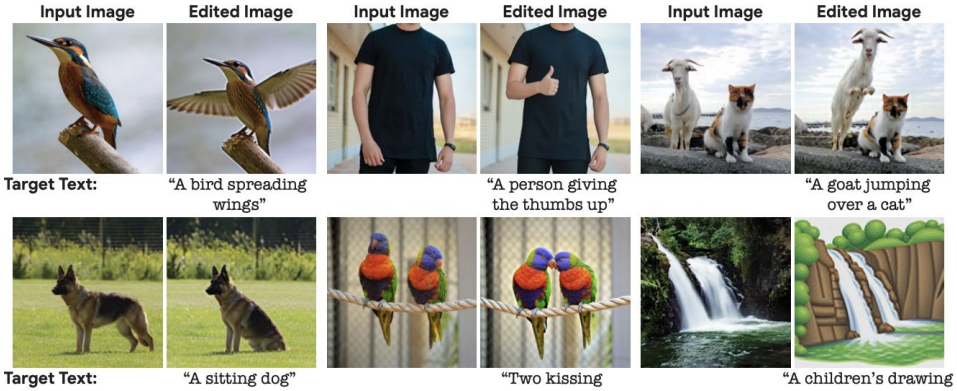
Personalization: fine-tune Text2Image DM for your object



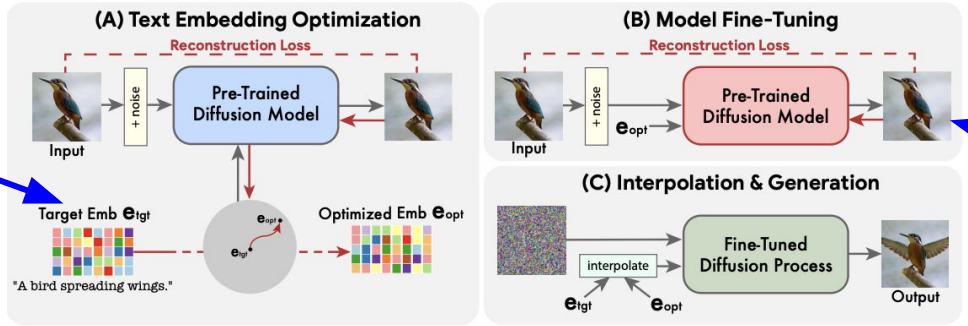
"Rare tokens" representing an object

Fine-tune T2I DM, through preserving semantic prior over the general object type (e.g dog)

Fine-tune LLM and Text2Image DM: Image Editing



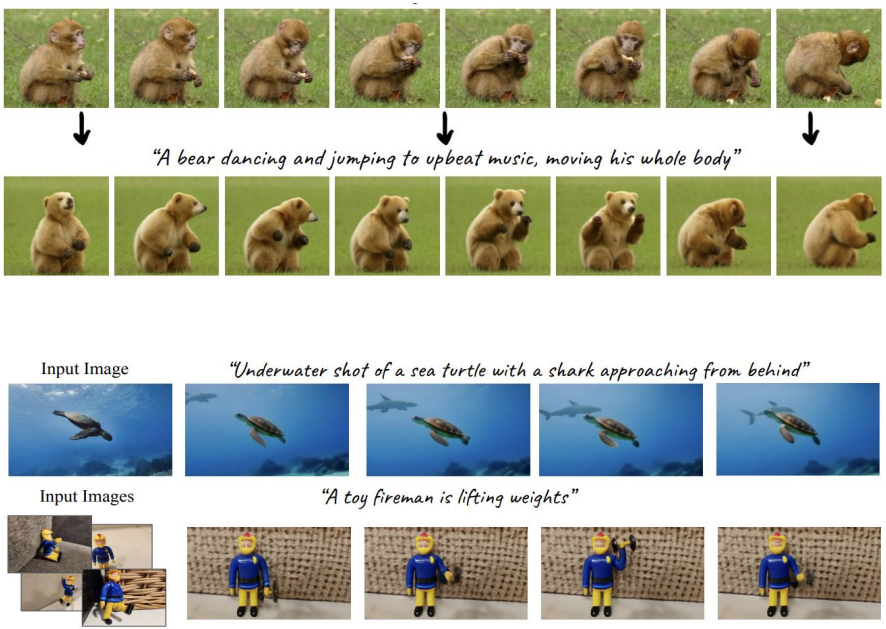
Fine-tune LLM



Fine-tune DM

Dreamix = ImagenVideo + DreamBooth: Video Editing and more

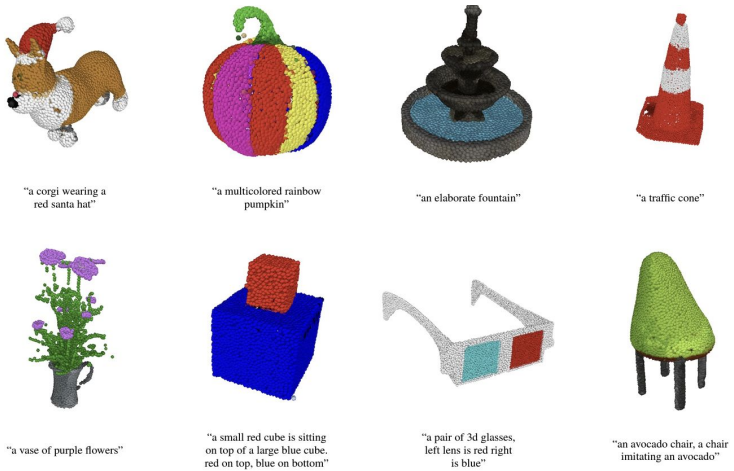
Fine-tune Text2Video Diffusion Model according to a textual description



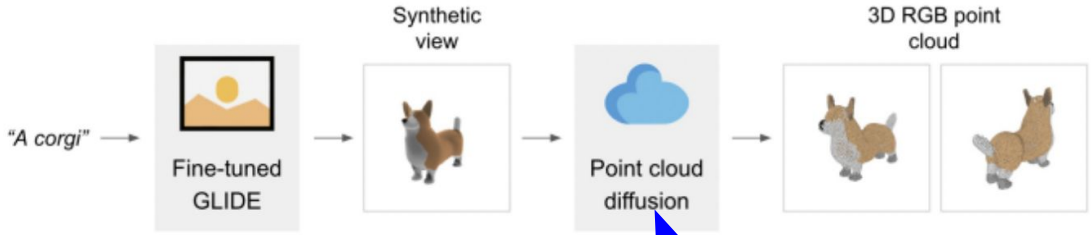
Fine-tuned T2I DM + I2PointCloud DM + SuperResolution DM

Point·E: A System for Generating 3D Point Clouds from Complex Prompts

Alex Nichol [*1] Heewoo Jun [*1] Prafulla Dhariwal [1] Pamela Mishkin [1] Mark Chen [1]



Generating Point Cloud with Text2Image DMs: Point-E



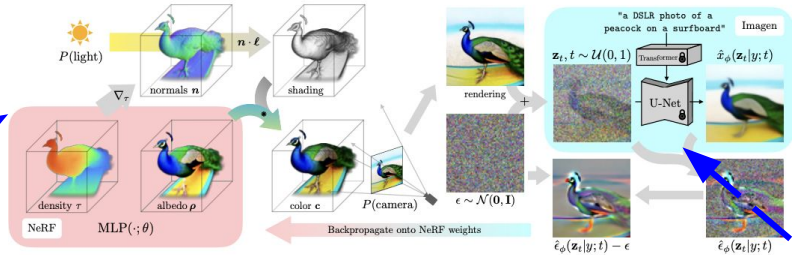
Fine-tuned Text2Image DM

Image-to-PointCloud DM
+ SuperResolution Point Cloud DM

T2I DM to evaluate/backprop errors of 3D NERF Model

DREAMFUSION: TEXT-TO-3D USING 2D DIFFUSION

Ben Poole[1], **Ajay Jain**[2], **Jonathan T. Barron**[1], **Ben Mildenhall**[1]
[1]Google Research, [2]UC Berkeley
{pooleb, barron, bmild}@google.com, ajayj@berkeley.edu



3D Model Parameterized with NERF(MLP): color, density

NERF MLP weights backprop

Frozen T2I Imagen DM: noise estimation

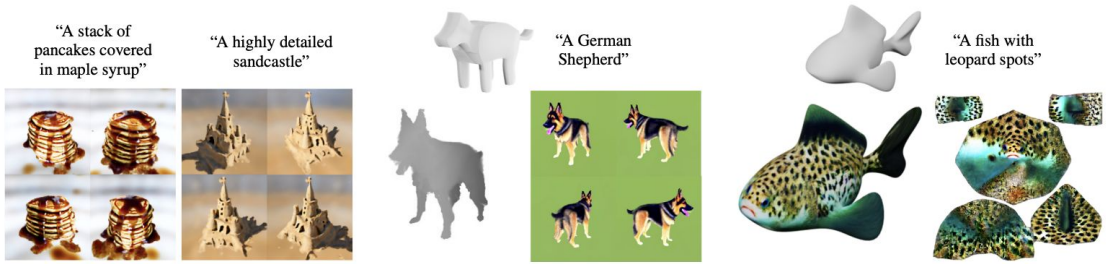
Latent DM to evaluate/backprop errors of 3D NERF Model

Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures

Gal Metzer* Elad Richardson* Or Patashnik Raja Giryes Daniel Cohen-Or

Tel Aviv University

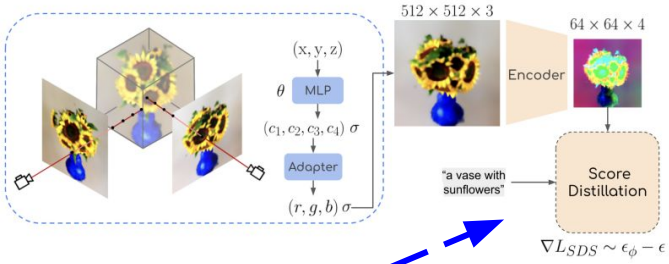
Build 3D model from its textual description accompanied with the shape sketch or its exact form



Latent-NeRF Sketch-Shape Latent-Paint

T2I DM to evaluate/backprop errors of 3D NERF Model

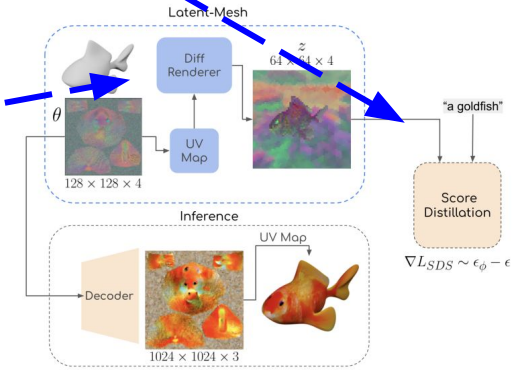
3D Latent NERF Model



NERF MLP weights backprop

Frozen Stable Diffusion (Latent): noise estimation

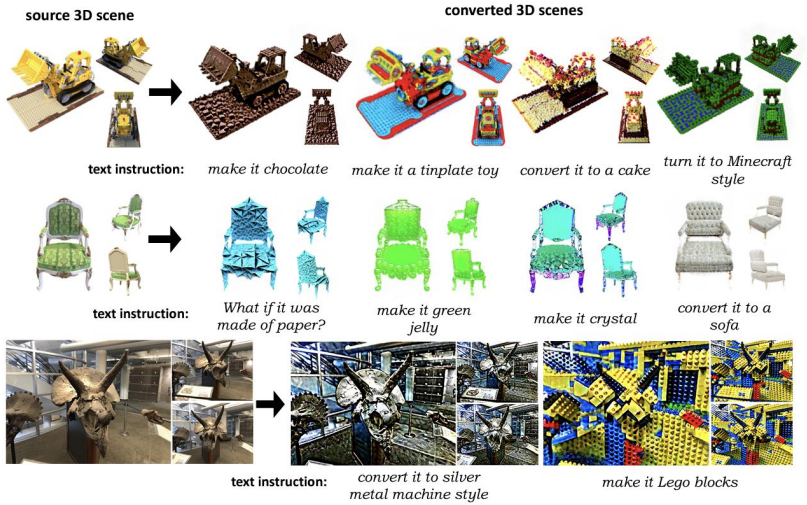
3D Latent NERF with UV texture generation



T2I DM to evaluate/backprop errors of 3D NERF Model 2

Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion

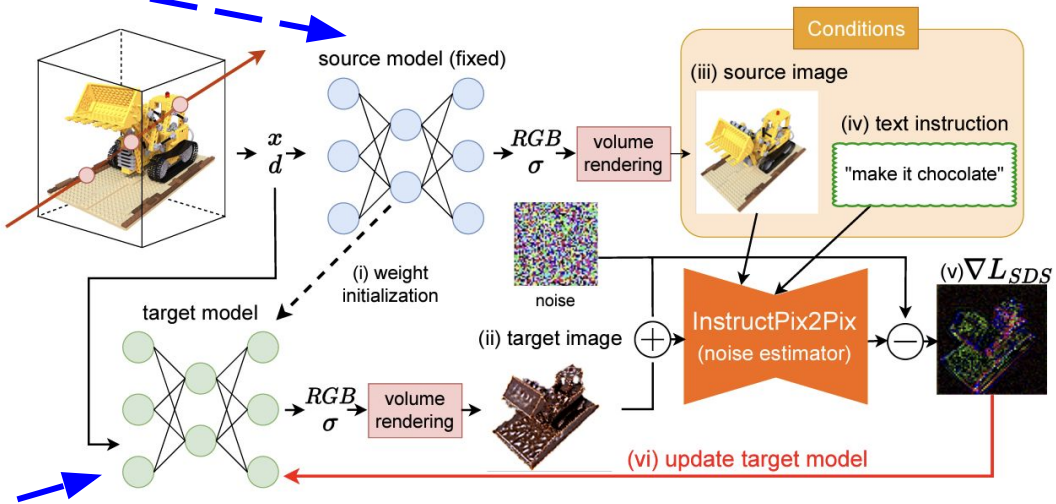
Hiromichi Kamata* Yuiko Sakuma Akio Hayakawa Masato Ishii Takuya Narihira
Sony Group Corporation
Tokyo, Japan



T2I DM to evaluate/backprop errors of 3D NERF Model, 2

Edit your 3D NERF Model with Text

3D NERF Model: Source



Frozen T2I Diffusion Model

3D NERF Model: Edited

Adapt DMs to receive input of different types (in addition to text)

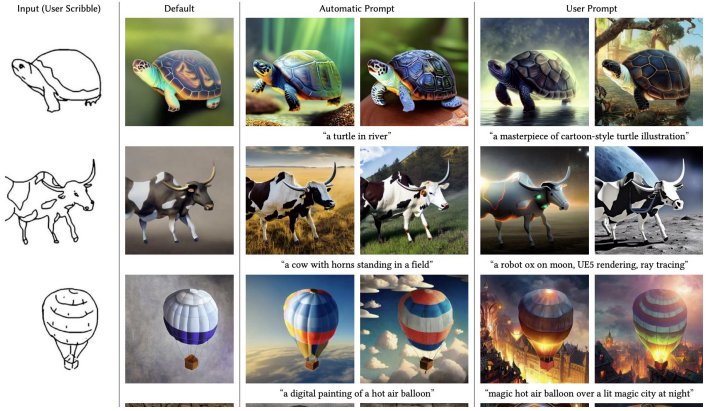
Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang and Maneesh Agrawala
Stanford University

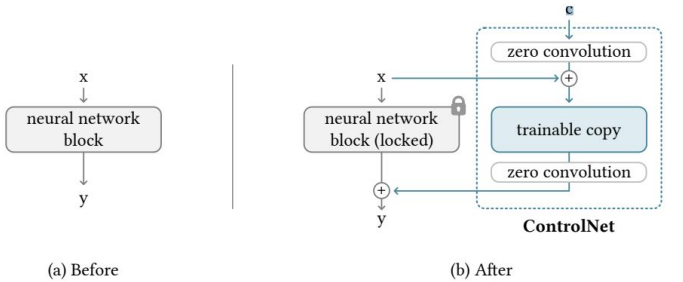
Condition your Text2Image Diffusion Dodel with:

- Canny Edges
- Hough Lines
- User Scribbles
- Holistically-Nested Edge(HED) Edges
- Segmentation Map
- Depth Map
- Pose
- ….

ControlNet: Control your DM with Almost Any Means



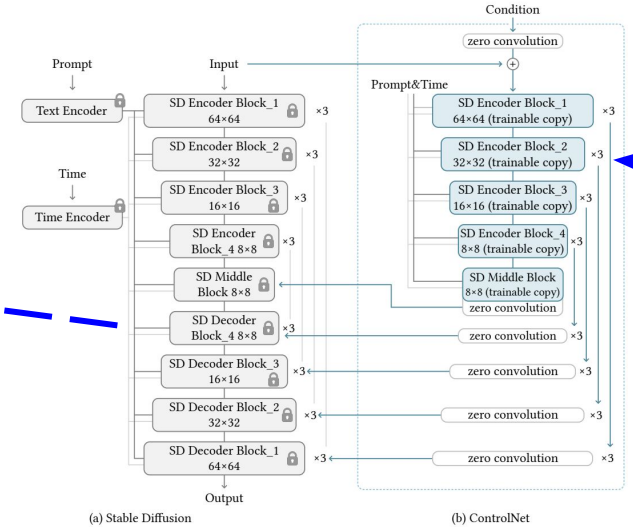
ControlNet: Control your DM with “Almost” Any Means



Train task-specific Text2Image DM to learn the conditional control

Combine this model with a frozen Text2Image DM to get “control” on DM

Frozen Text2Image DM



Combine **Frozen DM** With **Task-Specific DM** Through **Zero-Convolution**

What have we learnt today?

- What are Denoising Diffusion Probabilistic Models (DDPMs or just DMs)
- 2 Main Text2Image DMs Architectures
- 4 Broad Approaches to Utilize Text2Image DMs to create/edit many types of visual data
- Tasks for which we can harness Text2Image DMs?
  - Image Editing (Personalization)
  - Video Generation/Editing
  - 3D Models/Point Clouds Generation and Editing
  - Control Your Image with Scribble, Depth Maps, Hough Lines etc

Tiempo para Preguntas y Respuestas