Artificial Intelligence for Business Research @Antai

Unsupervised Learning: Diffusion Models

Renyu (Philip) Zhang

1

Agenda

- Denoised Diffusion Probabilistic Models
- Latent Diffusion, CLIP, DALL-E. Diffusion Transformer
- Potential Applications of Diffusions in Biz/Econ Research

2

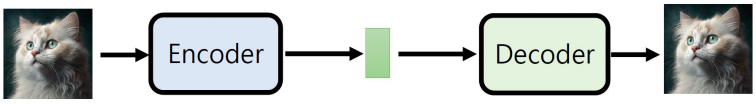
2

From VAE to Diffusions

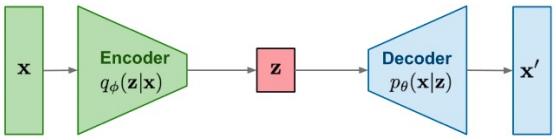
Denoising diffusion probabilistic models
J Ho, A Jain, P Abbeel - Advances in neural information ..., 2020 - proceedings.neurips.cc
... **Diffusion** models are straightforward to define and efficient to ... We show that **diffusion** models actually are capable of ... of **diffusion** models reveals an equivalence with **denoising** score ...
☆ Save 99 Cite Cited by 8273 Related articles All 9 versions

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

VAE

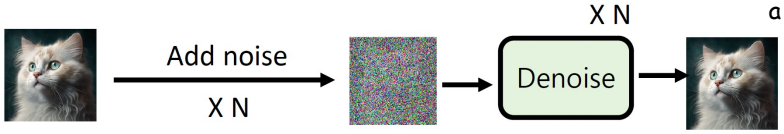


VAE: maximize variational lower bound



Diffusion models are incremental updates where the assembly of the whole gives us the encoder-decoder structure and the transition from one state to another is realized by a denoiser.

Diffusion



X N

Diffusion models: Gradually add Gaussian noise and then reverse

Denoised Diffusion Probabilistic Models (DDPM)

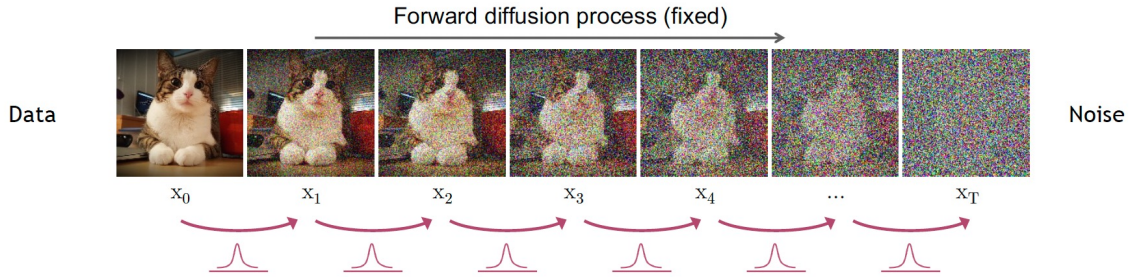
3

3

Forward Diffusion Process

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

The formal definition of the forward process in T steps:



$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad \Rightarrow \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \qquad \text{(joint)}$$

$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\varepsilon}_t$, where $\beta_t \in [0,1]$ and $\boldsymbol{\varepsilon}_t$ is d-dimensional standard normal independent of anything.

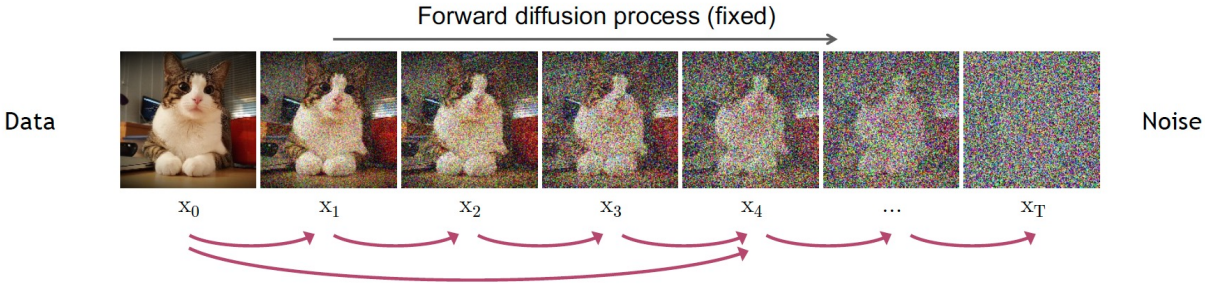
Also define $\alpha_t := 1 - \beta_t$.

4

4

Diffusion Kernel

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/



Define $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ ➡ $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}))$ (Diffusion Kernel)

For sampling: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\,\epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

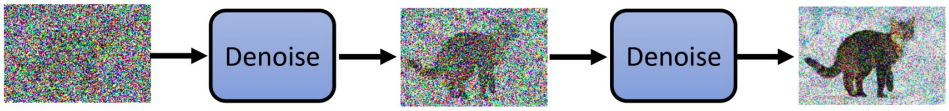
β_t values schedule (i.e., the noise schedule) is designed such that $\bar{\alpha}_T \to 0$ and $q(\mathbf{x}_T|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$

How do we reconstruct $\boldsymbol{x_0}$ from $\boldsymbol{x_T}$?

5

5

Denoising is like Sculpture



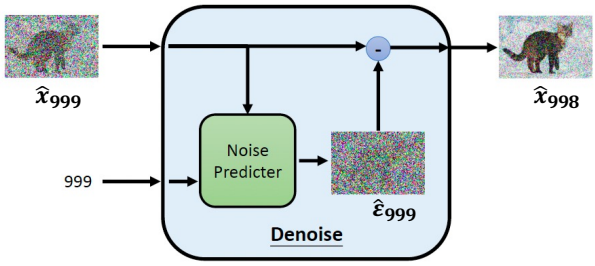
The sculpture is already complete within the marble block, before I start my work. It is already there, I just have to chisel away the superfluous material. **- Michelangelo**

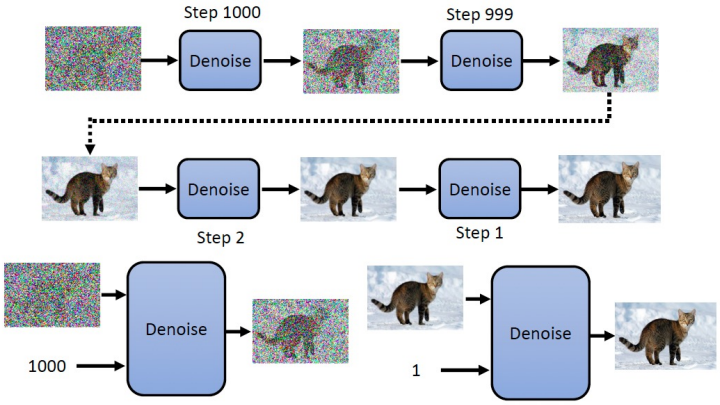
https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

6

6

Denoise





Gradually denoise a standard Gaussian x_T to generate the original data x_0.

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

7

7

Generative Learning with Denoising

Generation:

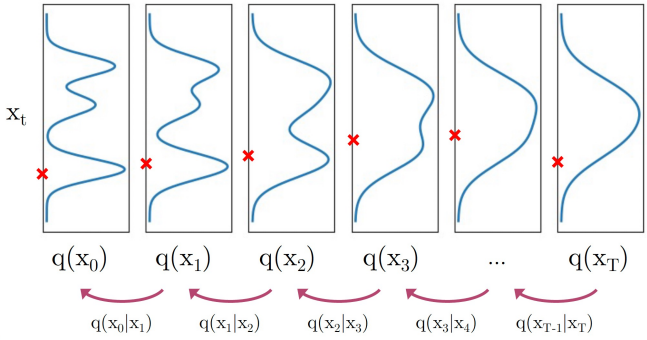
Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$

Iteratively sample $\mathbf{x}_{t-1} \sim \underbrace{q(\mathbf{x}_{t-1}|\mathbf{x}_t)}$

True Denoising Dist.

Equivalently, $q(\boldsymbol{\epsilon}_t|\boldsymbol{x}_t)$.

Diffused Data Distributions



In general, $q(\mathbf{x}_{t-1}|\mathbf{x}_t) \propto q(\mathbf{x}_{t-1})q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is intractable.

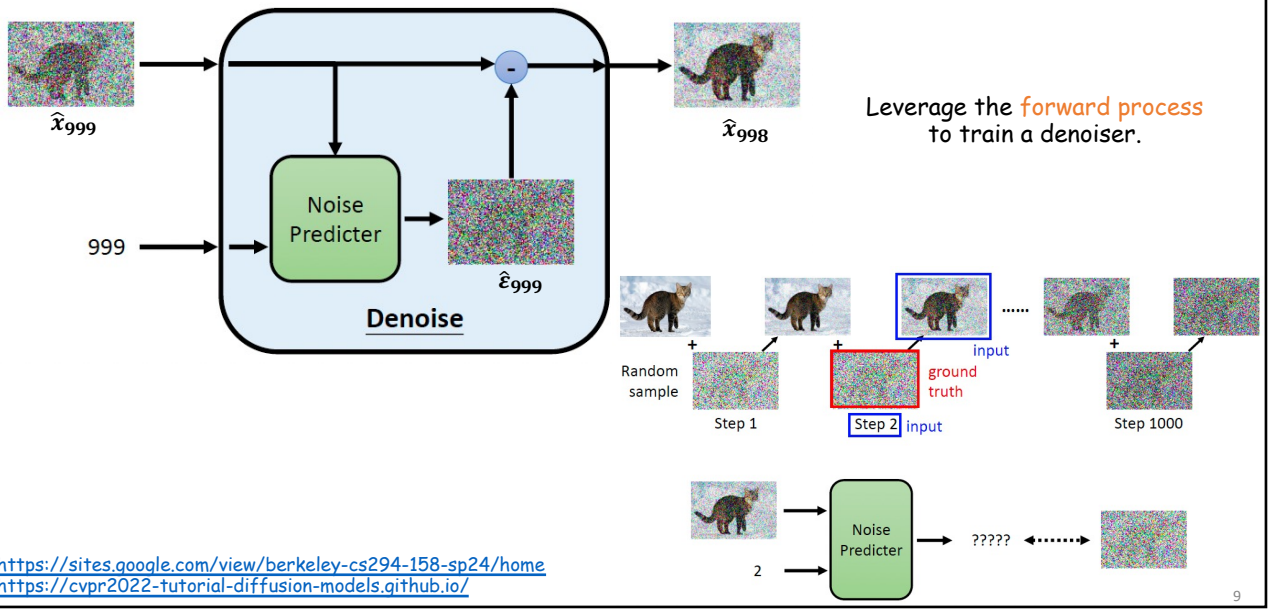
We can use a normal distribution to approximate $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$.

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

8

8

Training Denoisers

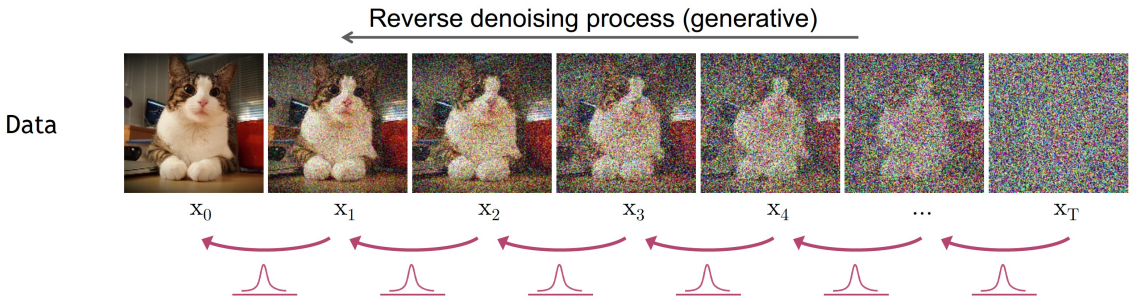


https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

9

9

Reverse Denoising Process



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}, \sigma_t^2\mathbf{I})$$

Trainable network
(U-net, Denoising Autoencoder)

$$\Rightarrow \quad p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

It can be derived that $\sigma_t^2 = \frac{(1-\alpha_t)(1-\alpha_{t-1})}{1-\alpha_t}$.

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

10

10

Evidence Lower Bound (ELBO)

$$\log\big(p(\boldsymbol{x})\big) = \log\big(p(\boldsymbol{x_0})\big) \geq \text{ELBO} = \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

- Reconstruction term: How good the neural nets $p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)$ is to recover $\boldsymbol{x}_0$ from $\boldsymbol{x}_1$ when the sample is drawn from $q(\boldsymbol{x}_1|\boldsymbol{x}_0)$.
- Prior matching term: How close the distribution of the final noisified input is to the standard Gaussian prior.
- Denoising matching term: How close the denoising transition distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ approximates the tractable ground-truth denoising transition step distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$.

It can be derived that $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I})$

- Train the noise $\widehat{\boldsymbol{\varepsilon}_\theta}(\boldsymbol{x}_t)$, which is a neural network that maps from $\boldsymbol{x}_t$ to $\boldsymbol{\varepsilon}_0$, where $\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\varepsilon}_0$

$$\text{ELBO}_{\boldsymbol{\theta}} = -\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}\left[\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2}\left\|\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_t) - \boldsymbol{\epsilon}_0\right\|^2\right].$$ Then, sample $\boldsymbol{x}_{t-1}$ from $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$.

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

11

11

DDPM Algorithms

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2$$
6: **until** converged

Algorithm 2 Sampling

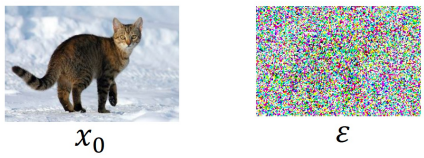
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

References: https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

12

12

DDPM Training

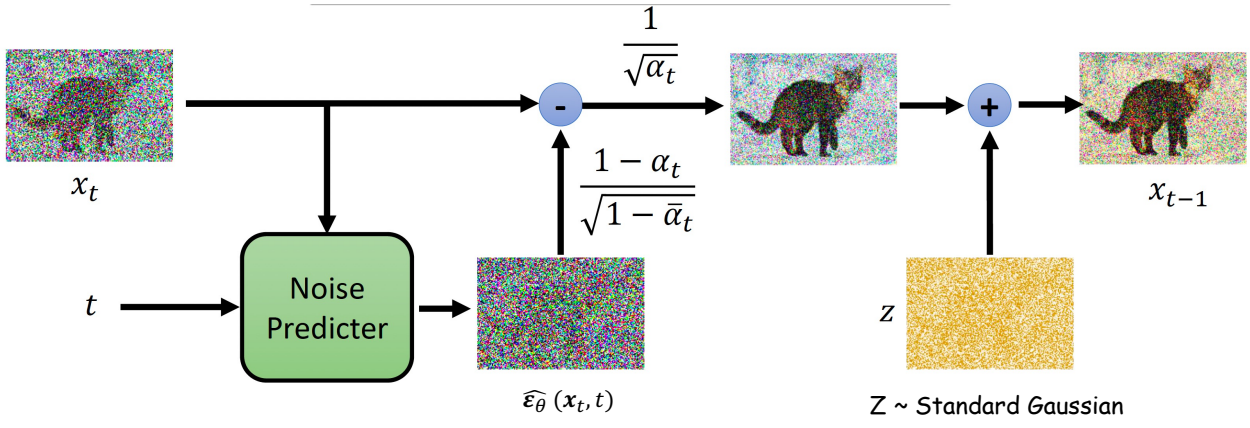


https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

13

13

DDPM Inference



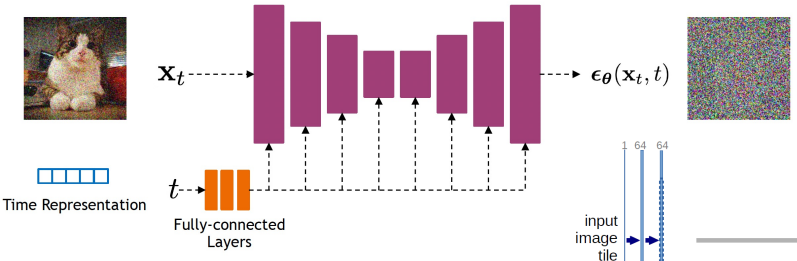
DDPM Inference is based on $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I})$

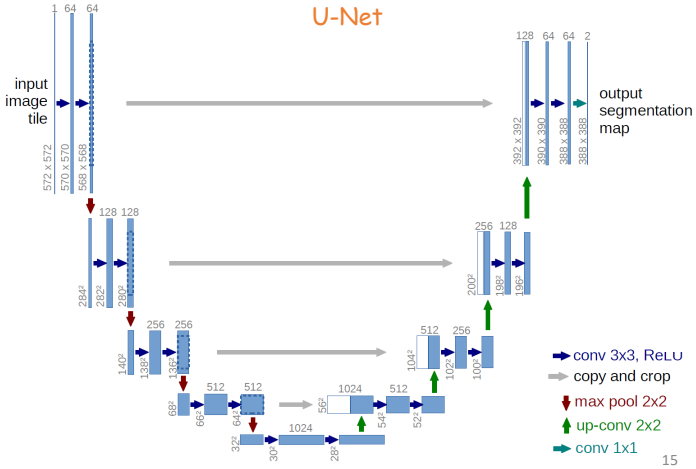
https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

14

14

Backbone: U-Net





Diffusion models often use U-Net architectures with ResNet blocks and self-attention layers to represent $\widehat{\varepsilon_\theta}(x_t, t)$.

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

15

15

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87±0.12	25.32	
SNGAN [39]	8.22±0.05	21.7	
SNGAN-DDLS [4]	9.09±0.10	15.42	
StyleGAN2 + ADA (v1) [29]	**9.74** ± 0.05	3.26	
Ours (L, fixed isotropic $\mathbf{\Sigma}$)	7.67±0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46±0.11	**3.17**	≤ 3.75 (3.72)

DDPM: Results

IS: Inception Score

FID: Fréchet Inception Distance

NLL Test: Negative Log-Likelihood Test

Figure 3: LSUN Church samples. FID=7.89

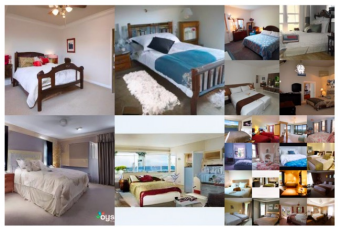


Figure 4: LSUN Bedroom samples. FID=4.90

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

16

16

Diffusion Models Beat GAN on Image Generation

Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Abstract

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 https://github.com/openai/guided-diffusion.

GANs [19] currently hold the state-of-the-art on most image generation tasks [5, 68, 28] as measured by sample quality metrics such as FID [23], Inception Score [54] and Precision [32]. However, some of these metrics do not fully capture diversity, and it has been shown that GANs capture less diversity than state-of-the-art likelihood-based models [51, 43, 42]. Furthermore, GANs are often difficult to train, collapsing without carefully selected hyperparameters and regularizers [5, 41, 4].

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

Diffusions are:

- Generating images of higher quality and diversity;
- Stabler to train;
- More flexible for control/conditioning (text-to-image);
- Theoretically grounded.

17

17

Agenda

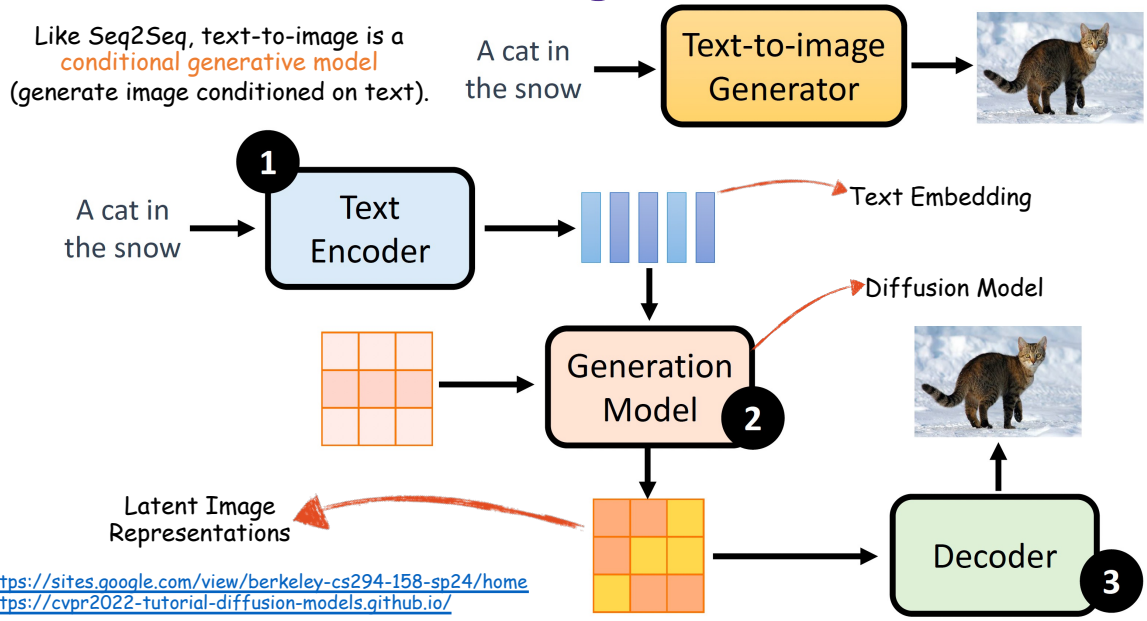
- Denoised Diffusion Probabilistic Models
- Latent Diffusion, CLIP, DALL-E. Diffusion Transformer
- Potential Applications of Diffusions in Biz/Econ Research

18

18

Text-to-Image Generation

Like Seq2Seq, text-to-image is a conditional generative model (generate image conditioned on text).

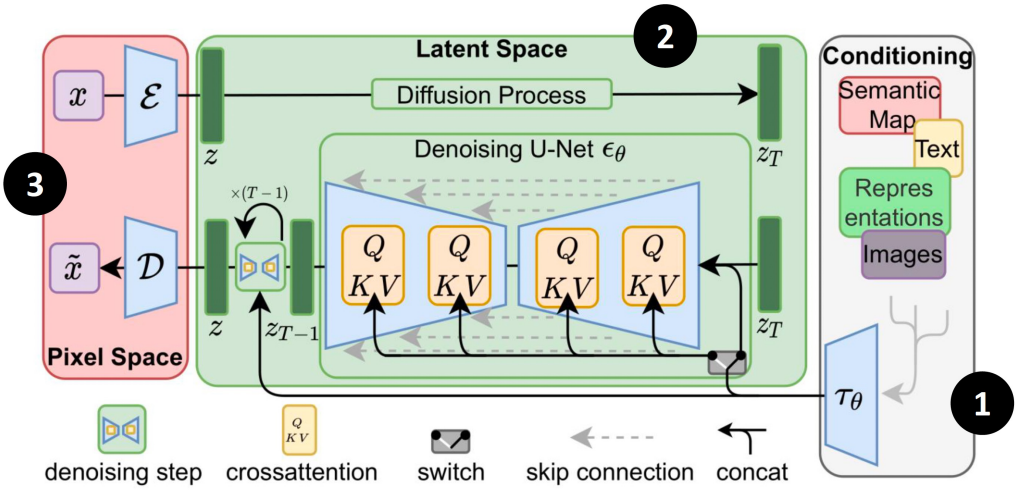


https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

19

19

Stable/Latent Diffusion Model



High-resolution image synthesis with latent diffusion models
R Rombach, A Blattmann, D Lorenz… - Proceedings of the …, 2022 - openaccess.thecvf.com
… To lower the computational demands of training diffusion **models** towards **high-resolution image synthesis**, we observe that although **diffusion models** allow to ignore perceptually …
☆ Save 99 Cite Cited by 6858 Related articles All 12 versions »

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

20

20

Stable Diffusion Results

Text-to-Image Synthesis on LAION. 1.45B Model.

Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

High-resolution image synthesis with latent diffusion models

R Rombach, A Blattmann, D Lorenz… - Proceedings of the …, 2022 - openaccess.thecvf.com

… To lower the computational demands of training diffusion **models** towards **high-resolution image synthesis**, we observe that although **diffusion models** allow to ignore perceptually …

☆ Save 99 Cite Cited by 6858 Related articles All 12 versions »

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

CelebA-HQ 256 × 256			
Method	**FID ↓**	**Prec. ↑**	**Recall ↑**
DC-VAE [63]	15.8	-	-
VQGAN+T [23] (k=400)	10.2	-	-
PGGAN [39]	8.0	-	-
LSGM [93]	7.22	-	-
UDM [43]	7.16	-	-
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49

FFHQ 256 × 256			
Method	**FID ↓**	**Prec. ↑**	**Recall ↑**
ImageBART [21]	9.57	-	-
U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
UDM [43]	5.54	-	-
StyleGAN [41]	4.16	0.71	0.46
ProjectedGAN [76]	3.08	0.65	0.46
LDM-4 (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256			
Method	**FID ↓**	**Prec. ↑**	**Recall ↑**
DDPM [30]	7.89	-	-
ImageBART [21]	7.32	-	-
PGGAN [39]	6.42	-	-
StyleGAN [41]	4.21	-	-
StyleGAN2 [42]	3.86	-	-
ProjectedGAN [76]	**1.59**	0.61	0.44
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**

LSUN-Bedrooms 256 × 256			
Method	**FID ↓**	**Prec. ↑**	**Recall ↑**
ImageBART [21]	5.51	-	-
DDPM [30]	4.9	-	-
UDM [43]	4.57	-	-
StyleGAN [41]	2.35	0.59	0.48
ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.52**	0.61	0.34
LDM-4 (ours, 200-s)	2.95	**0.66**	0.48

Table 1. Evaluation metrics for unconditional image synthesis. CelebA-HQ results reproduced from [43, 63, 100], FFHQ from [42, 43]. †: N-s refers to N sampling steps with the DDIM [84] sampler. *: trained in *KL*-regularized latent space. Additional results can be found in the supplementary.

Text-Conditional Image Synthesis				
Method	**FID ↓**	**IS↑**	N_{params}	
CogView† [17]	27.10	18.20	4B	self-ranking, rejection rate 0.017
LAFITE† [109]	26.94	26.02	75M	
GLIDE* [59]	12.24	-	6B	277 DDIM steps, c.f.g. [32] $s = 3$
Make-A-Scene* [26]	**11.84**	-	4B	c.f.g for AR models [98] $s = 5$
LDM-KL-8	23.31	$20.03^{\pm 0.33}$	1.45B	250 DDIM steps
*LDM-KL-8-G**	12.63	$\mathbf{30.29}^{\pm 0.42}$	1.45B	250 DDIM steps, c.f.g. [32] $s = 1.5$

Table 2. Evaluation of text-conditional image synthesis on the 256 × 256-sized MS-COCO [51] dataset: with 250 DDIM [84] steps our model is on par with the most recent diffusion [59] and autoregressive [26] methods despite using significantly less parameters. †/*:Numbers from [109]/ [26]

21

21

Contrastive Language-Image Pre-training (CLIP)

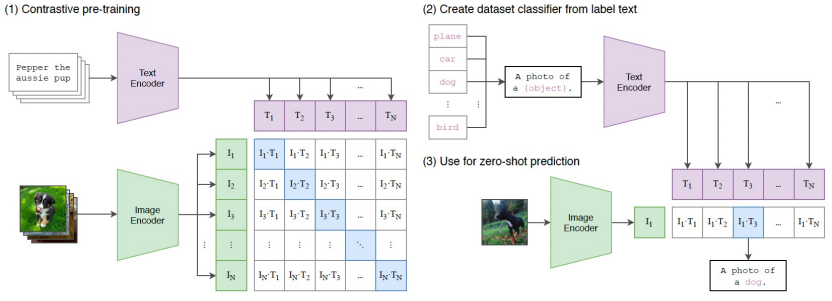
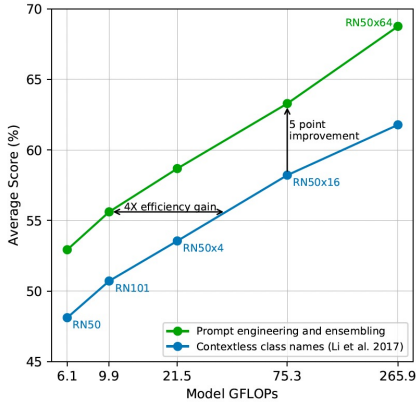


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.



https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

Learning transferable visual models from natural language supervision

A Radford, JW Kim, C Hallacy… - … machine learning, 2021 - proceedings.mlr.press

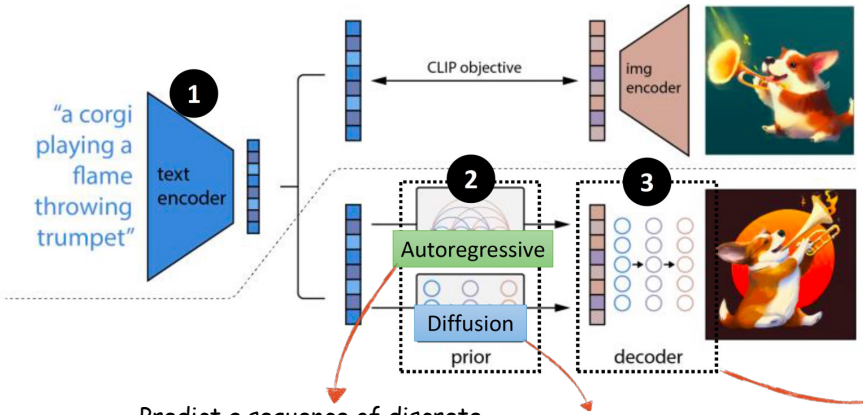
… We speculate this is due to natural language providing wider supervision for visual concepts involving verbs, compared to the noun-centric object supervision in ImageNet. …

☆ 保存 99 引用 被引用次数：14959 相关文章 所有 19 个版本 »

22

22

DALL·E 2



Model	FID	Zero-shot FID	Zero-shot FID (filt)
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**		
DALL-E (Ramesh et al., 2021)		~ 28	
LAFITE (Zhou et al., 2021)		26.94	
GLIDE (Nichol et al., 2021)		12.24	12.89
Make-A-Scene (Gafni et al., 2022)			11.84
unCLIP (AR prior)		10.63	11.08
unCLIP (Diffusion prior)		**10.39**	**10.87**

MS-COCO Dataset

Cascaded diffusion models: 64x64 → 256x256 → 1024x1024

Predict a sequence of discrete codes auto-regressively.

Model the continuous image embedding by diffusion models conditioned on caption

Produce images conditioned on CLIP embeddings (and text).

Zero-shot text-to-image generation

A Ramesh, M Pavlov, G Goh, S Gray… - International …, 2021 - proceedings.mlr.press

… We evaluate our model **zero-shot** by comparing it to three prior … our model to be capable of **zero-shot** image-to-image translation … We provide additional examples of **zero-shot** image-to-…

☆ Save 99 Cite Cited by 3553 Related articles All 5 versions »

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

[PDF] **Hierarchical text-conditional image generation with clip latents**

A Ramesh, P Dhariwal, A Nichol, C Chu… - arXiv preprint arXiv …, 2022 - 3dvar.com

… of **text-conditional image generation**. We first train a diffusion decoder to invert the **CLIP image** … , and can produce multiple **images** corresponding to a given **image** embedding. The …

☆ Save 99 Cite Cited by 4115 Related articles All 3 versions »

23

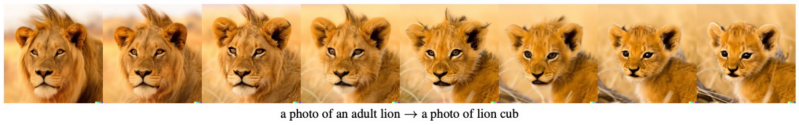
23

DALL·E 2

Text Diffs

a photo of a cat → an anime drawing of a super saiyan cat, artstation

a photo of a victorian house → a photo of a modern house



a photo of an adult lion → a photo of lion cub

Change the image CLIP embedding towards the difference of the text CLIP embeddings of two prompts.

Decoder latent is kept as a constant.

https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/

[PDF] **Hierarchical text-conditional image generation with clip latents**

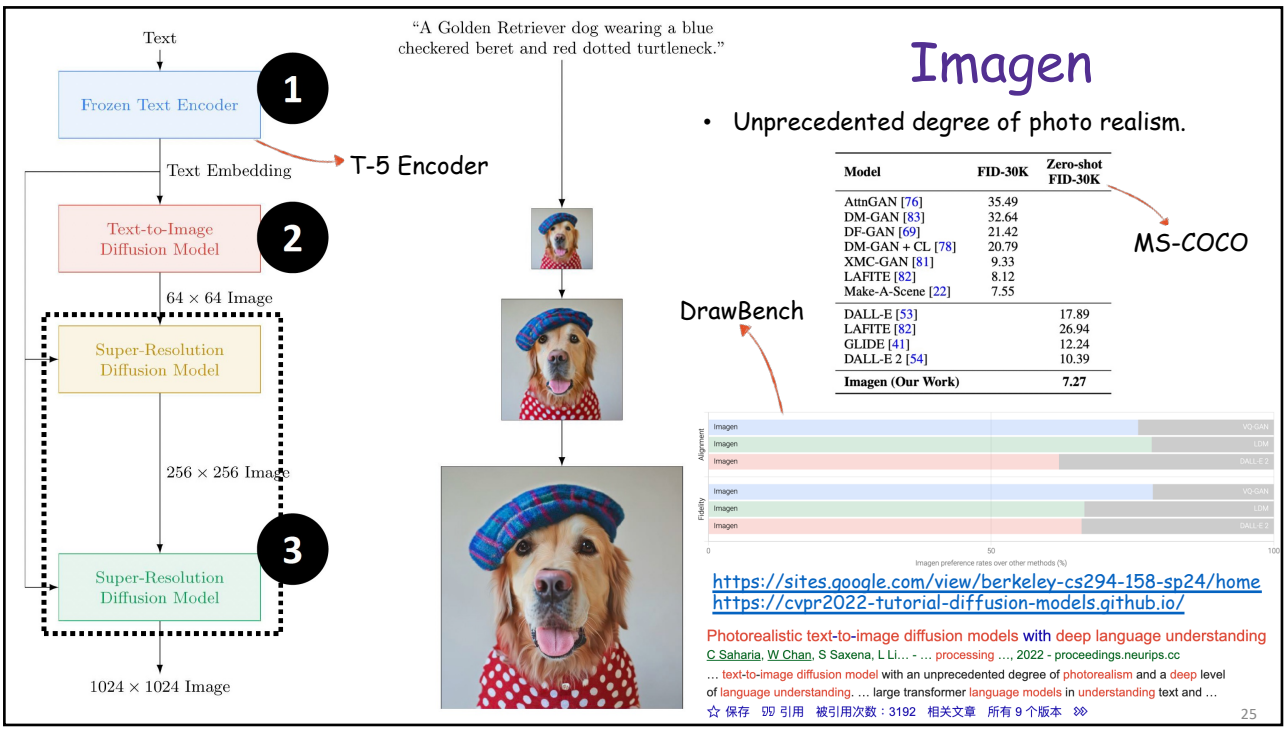
A Ramesh, P Dhariwal, A Nichol, C Chu… - arXiv preprint arXiv …, 2022 - 3dvar.com

… of **text-conditional image generation**. We first train a diffusion decoder to invert the **CLIP image** … , and can produce multiple **images** corresponding to a given **image** embedding. The …

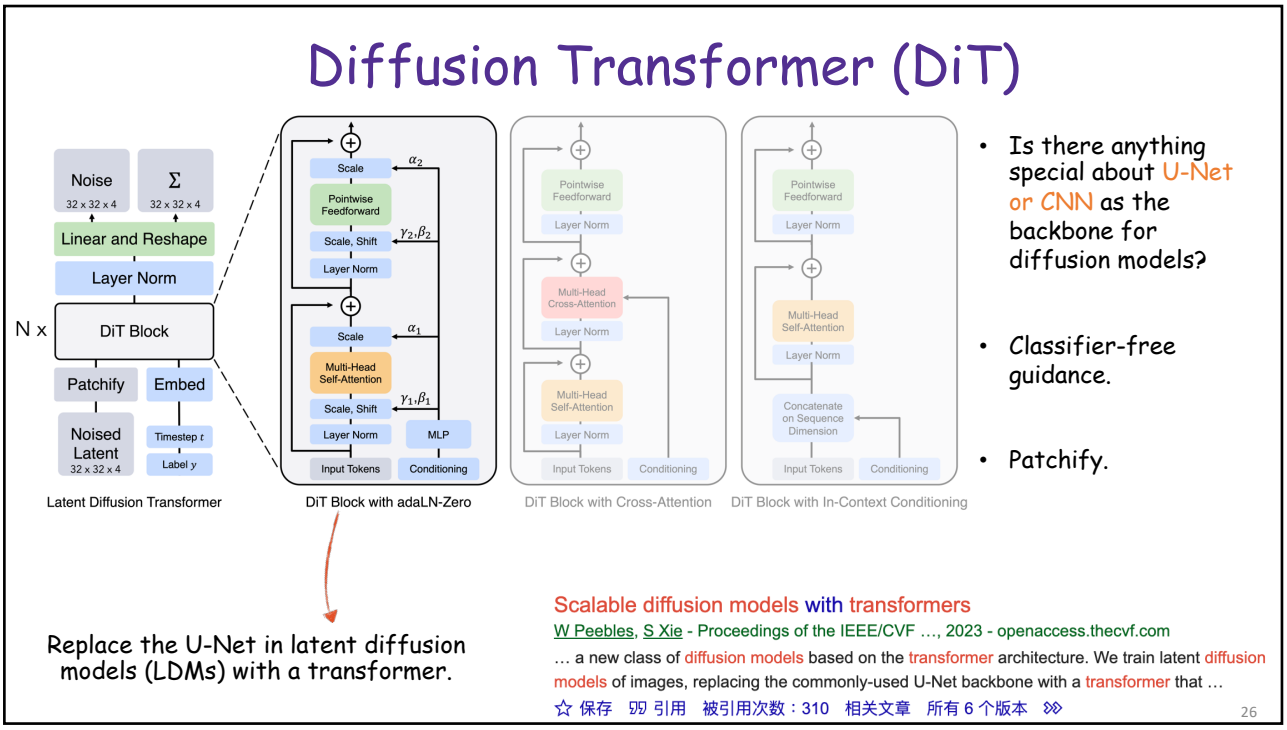
☆ Save 99 Cite Cited by 4115 Related articles All 3 versions »

24

24

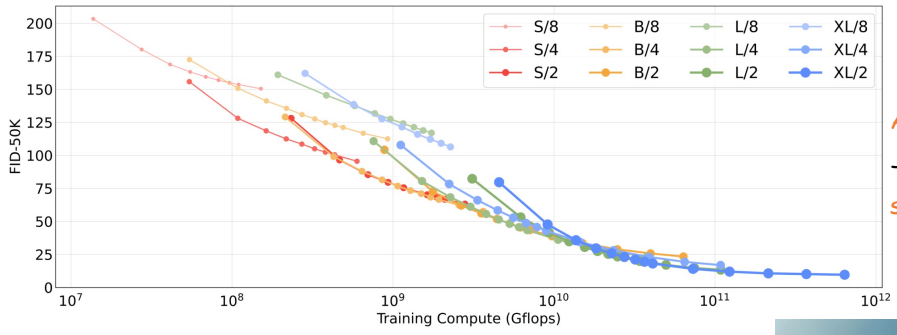

Text
Frozen Text Encoder
1
T-5 Encoder
Text Embedding
Text-to-Image Diffusion Model
2
64 × 64 Image
Super-Resolution Diffusion Model
256 × 256 Image
3
Super-Resolution Diffusion Model
1024 × 1024 Image
"A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck."
Imagen
• Unprecedented degree of photo realism.
Model
FID-30K
Zero-shot FID-30K
AttnGAN [76] 35.49
DM-GAN [83] 32.64
DF-GAN [69] 21.42
DM-GAN + CL [78] 20.79
XMC-GAN [81] 9.33
LAFITE [82] 8.12
Make-A-Scene [22] 7.55
DALL-E [53] 17.89
LAFITE [82] 26.94
GLIDE [41] 12.24
DALL-E 2 [54] 10.39
Imagen (Our Work) 7.27
MS-COCO
DrawBench
Alignment
Fidelity
Imagen preference rates over other methods (%)
https://sites.google.com/view/berkeley-cs294-158-sp24/home
https://cvpr2022-tutorial-diffusion-models.github.io/
Photorealistic text-to-image diffusion models with deep language understanding
C Saharia, W Chan, S Saxena, L Li... - ... processing ..., 2022 - proceedings.neurips.cc
... text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. ... large transformer language models in understanding text and ...
☆ 保存 99 引用 被引用次数：3192 相关文章 所有 9 个版本 »
25

25


Diffusion Transformer (DiT)
Noise
32 x 32 x 4
Σ
32 x 32 x 4
Linear and Reshape
Layer Norm
N x
DiT Block
Patchify
Embed
Noised Latent
32 x 32 x 4
Timestep t
Label y
Latent Diffusion Transformer
Scale
Pointwise Feedforward
Scale, Shift
Layer Norm
Multi-Head Self-Attention
MLP
Input Tokens
Conditioning
DiT Block with adaLN-Zero
Multi-Head Cross-Attention
DiT Block with Cross-Attention
Concatenate on Sequence Dimension
DiT Block with In-Context Conditioning
• Is there anything special about U-Net or CNN as the backbone for diffusion models?
• Classifier-free guidance.
• Patchify.
Replace the U-Net in latent diffusion models (LDMs) with a transformer.
Scalable diffusion models with transformers
W Peebles, S Xie - Proceedings of the IEEE/CVF ..., 2023 - openaccess.thecvf.com
... a new class of diffusion models based on the transformer architecture. We train latent diffusion models of images, replacing the commonly-used U-Net backbone with a transformer that ...
☆ 保存 99 引用 被引用次数：310 相关文章 所有 6 个版本 »
26

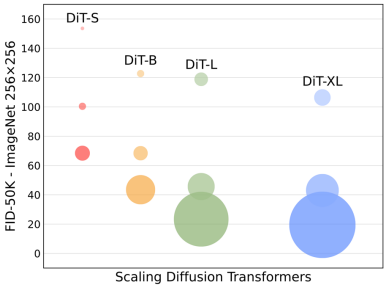
26

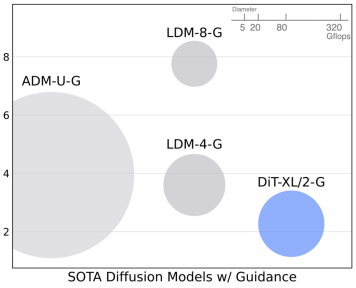
Diffusion Transformer (DiT) Performances



Attention is all you need again!

The bitter/sweet lesson: scaling laws.





27

27

Video Generation Models Scale

Base Compute

4x Compute

32x Compute

https://openai.com/research/video-generation-models-as-world-simulators

28

28

Agenda

- Denoised Diffusion Probabilistic Models
- Latent Diffusion, CLIP, DALL-E. Diffusion Transformers
- Potential Applications of Diffusions in Biz/Econ Research

29

29

ML-Driven Hypothesis Generation

Machine learning as a tool for hypothesis generation
J Ludwig, S Mullainathan
2023 · nber.org

Abstract
While hypothesis testing is a highly formalized activity, hypothesis generation remains largely informal. We propose a systematic procedure to generate novel hypotheses about human behavior, which uses the capacity of machine learning algorithms to notice patterns people might not. We illustrate the procedure with a concrete application: judge decisions about who to jail. We begin with a striking fact: The defendant's face alone matters greatly for the judge's jailing decision. In fact, an algorithm given only the pixels in the defendant's mugshot accounts for up to half of the predictable variation. We develop a procedure that allows human subjects to interact with this black-box algorithm to produce hypotheses about what in the face influences judge decisions. The procedure generates hypotheses that are both interpretable and novel: They are not explained by demographics (eg race) or existing psychology research; nor are they already known (even if tacitly) to people or even experts. Though these results are specific, our procedure is general. It provides a way to produce novel, interpretable hypotheses from any highdimensional dataset (eg cell phones, satellites, online behavior, news headlines, corporate filings, and high-frequency time series). A central tenet of our paper is that hypothesis generation is in and of itself a valuable activity, and hope this encourages future work in this largely "prescientific" stage of science.

nber.org

SHOW LESS

☆ Save 99 Cite Cited by 17 Related articles All 11 versions »

- Directly use CNN algorithms to generate interpretable and testable hypotheses on human behaviors.
- Mug shots + ML predicts judge behavior in jailing decisions, uncovering (about 22.3%) new information never discovered before.
- Create counterfactual mug shots based on the algorithmic discovery and iterate with crowd-sourced workers to confirm what the human judges see is the same as what the ML algorithm sees, thus formalizing the algorithmic discoveries as testable hypotheses.

Can we do this part better with diffusion models?

30

30

Product aesthetic design: A machine learning augmentation

A Burnap, JR Hauser, A Timoshenko

Marketing Science, 2023 · pubsonline.informs.org

Aesthetics are critically important to market acceptance. In the automotive industry, an improved aesthetic design can boost sales by 30% or more. Firms invest heavily in designing and testing aesthetics. A single automotive "theme clinic" can cost more than $100,000, and hundreds are conducted annually. We propose a model to augment the commonly used aesthetic design process by predicting aesthetic scores and automatically generating innovative and appealing product designs. The model combines a probabilistic variational autoencoder (VAE) with adversarial components from generative adversarial networks (GAN) and a supervised learning component. We train and evaluate the model with data from an automotive partner—images of 203 SUVs evaluated by targeted consumers and 180,000 high-quality unrated images. Our model predicts well the appeal of new aesthetic designs—43.5% improvement relative to a uniform baseline and substantial improvement over conventional machine learning models and pretrained deep neural networks. New automotive designs are generated in a controllable manner for use by design teams. We empirically verify that automatically generated designs are (1) appealing to consumers and (2) resemble designs that were introduced to the market five years after our data were collected. We provide an additional proof-of-concept application using open-source images of dining room chairs.

History: Puneet Manchanda served as the senior editor.

Funding: A. Burnap received support from General Motors to partially fund a postdoctoral research position for the research conducted in this work. He certifies that none of the research or its results were censored or obfuscated in its publication. J. Hauser and A. Timoshenko certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Supplemental Material: The data files are available at https://doi.org/10.1287/mksc.2022.1429.

INFORMS

SHOW LESS

Save Cite Cited by 20 Related articles All 11 versions Web of Science: 4

Product Aesthetics Design

- Use VAE + GAN to improve product aesthetics design.
- Better predicts the appeal of new aesthetics designs compared with benchmarks.
- Automatically generated designs are (1) appealing to consumers and (2) resemble designs that were introduced to the market.

Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design

R Dew, A Ansari, O Toubia

Marketing Science, 2022 · pubsonline.informs.org

Logos serve a fundamental role as the visual figureheads of brands. Yet, because of the difficulty of using unstructured image data, prior research on logo design has largely been limited to nonquantitative studies. In this work, we explore the interplay between logo design and brand identity creation from a data-driven perspective. We develop both a novel logo feature extraction algorithm that uses modern image processing tools to decompose pixel-level image data into meaningful features and a multiview representation learning framework that links these visual features to textual descriptions, consumer ratings of brand personality, and other high-level tags describing firms. We apply this framework to a unique data set of brands to understand which brands use which logo features and how consumers evaluate these brands' personalities. Moreover, we show that manipulating the model's learned representations through what we term "brand arithmetic" yields new brand identities and can help with ideation. Finally, through an application to fast-food branding, we show how our model can be used as a decision support tool for suggesting typical logo features for a brand and for predicting consumers' reactions to new brands or rebranding efforts.

INFORMS

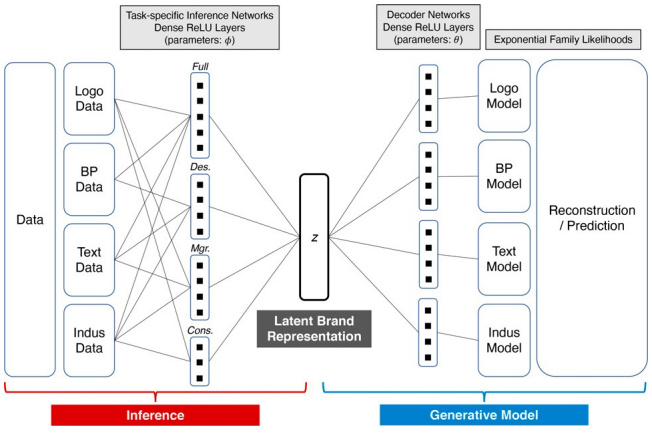
SHOW LESS

Save Cite Cited by 58 Related articles All 7 versions Web of Science: 6

1. Can we generate logos (or other creatives) in a more controlled fashion with diffusion models?
2. Can we use LLM as consumer simulators to evaluate the logos by GenAI?

GenAI for Logo Design

- Apply image processing/segmentation algorithms to extract information from brand logos.
- Apply multimodal-VAE the learn the latent representation of logos, and to generate new logos and predict consumer preferences.



Impact of GenAI on Art Creativity

Generative artificial intelligence, human creativity, and art

E Zhou, D Lee

PNAS nexus, 2024 · academic.oup.com

Abstract

Recent artificial intelligence (AI) tools have demonstrated the ability to produce outputs traditionally considered creative. One such system is text-to-image generative AI (e.g. Midjourney, Stable Diffusion, DALL-E), which automates humans' artistic execution to generate digital artworks. Utilizing a dataset of over 4 million artworks from more than 50,000 unique users, our research shows that over time, text-to-image AI significantly enhances human creative productivity by 25% and increases the value as measured by the likelihood of receiving a favorite per view by 50%. While peak artwork Content Novelty, defined as focal subject matter and relations, increases over time, average Content Novelty declines, suggesting an expanding but inefficient idea space. Additionally, there is a consistent reduction in both peak and average Visual Novelty, captured by pixel-level stylistic elements. Importantly, AI-assisted artists who can successfully explore more novel ideas, regardless of their prior originality, may produce artworks that their peers evaluate more favorably. Lastly, AI adoption decreased value capture (favorites earned) concentration among adopters. The results suggest that ideation and filtering are likely necessary skills in the text-to-image process, thus giving rise to "generative synesthesia"—the harmonious blending of human exploration and AI exploitation to discover new creative workflows.

Oxford University Press

SHOW LESS

Save Cite Related articles All 6 versions

- Text-to-image AI significant increases human creative work productivity by 25%, the likelihood of receiving review by 50%.
- "Creativity" decreases.
- Human exploration + AI exploitation.

33

33