

DSME 6635: Artificial Intelligence for Business Research

Unsupervised Learning: Diffusion Models

Renyu (Philip) Zhang

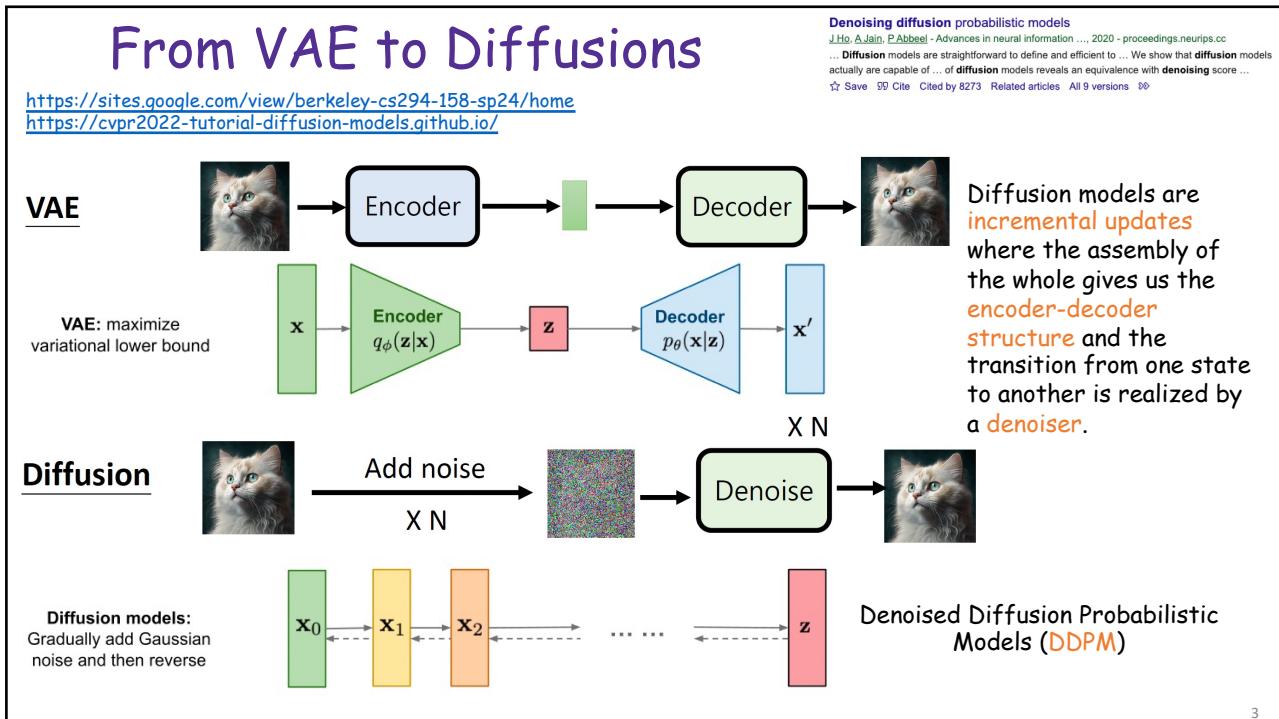
1

Agenda

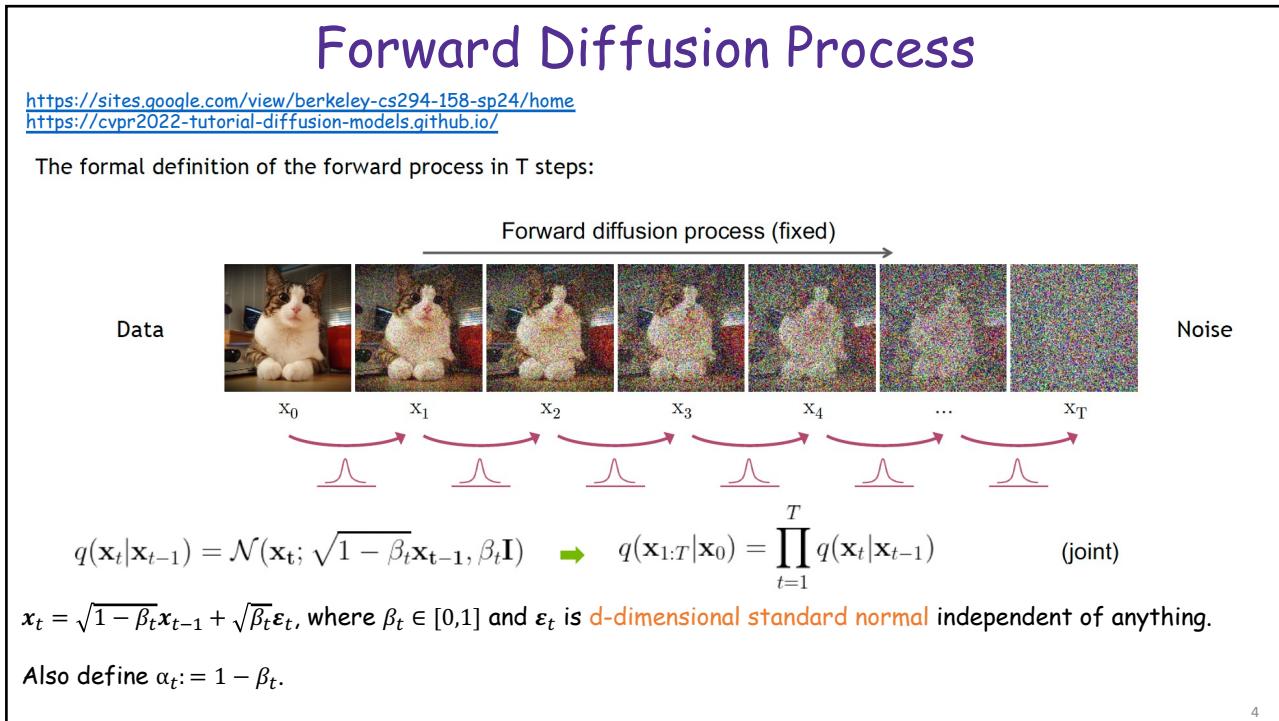
- Denoised Diffusion Probabilistic Models
- Latent Diffusion, CLIP, DALL-E. Diffusion Transformer
- Potential Applications of Diffusions in Biz/Econ Research

2

2



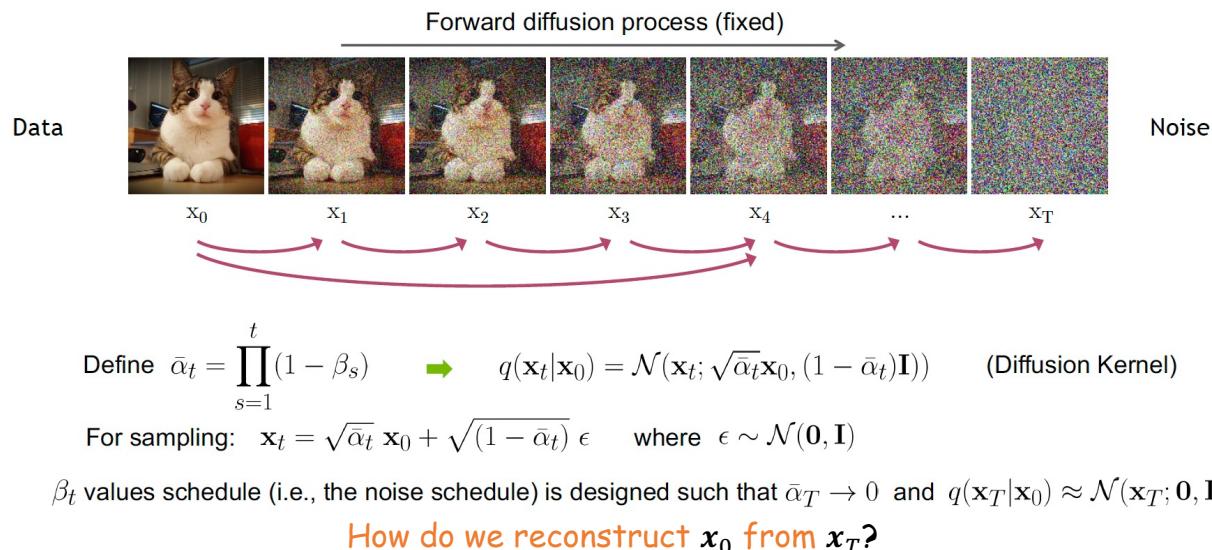
3



4

Diffusion Kernel

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



5

5

Denoising is like Sculpture



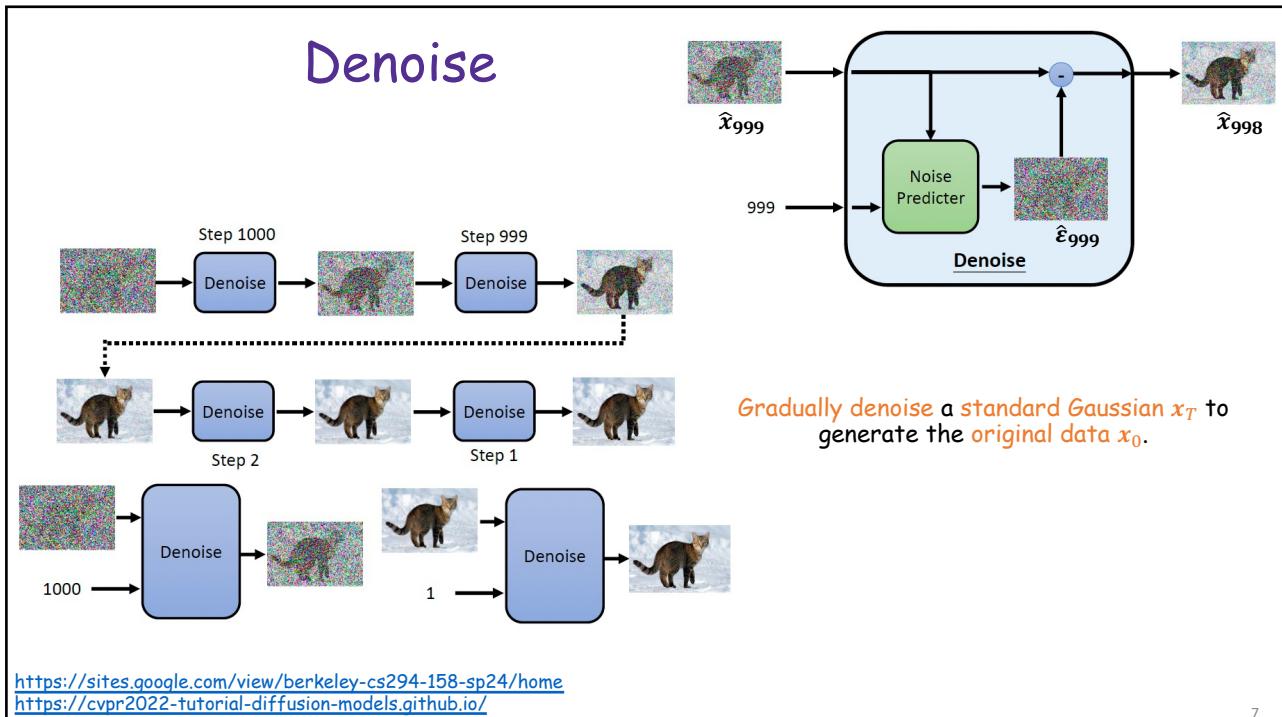
The sculpture is already complete within the marble block, before I start my work. It is already there, I just have to chisel away the superfluous material. - Michelangelo



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

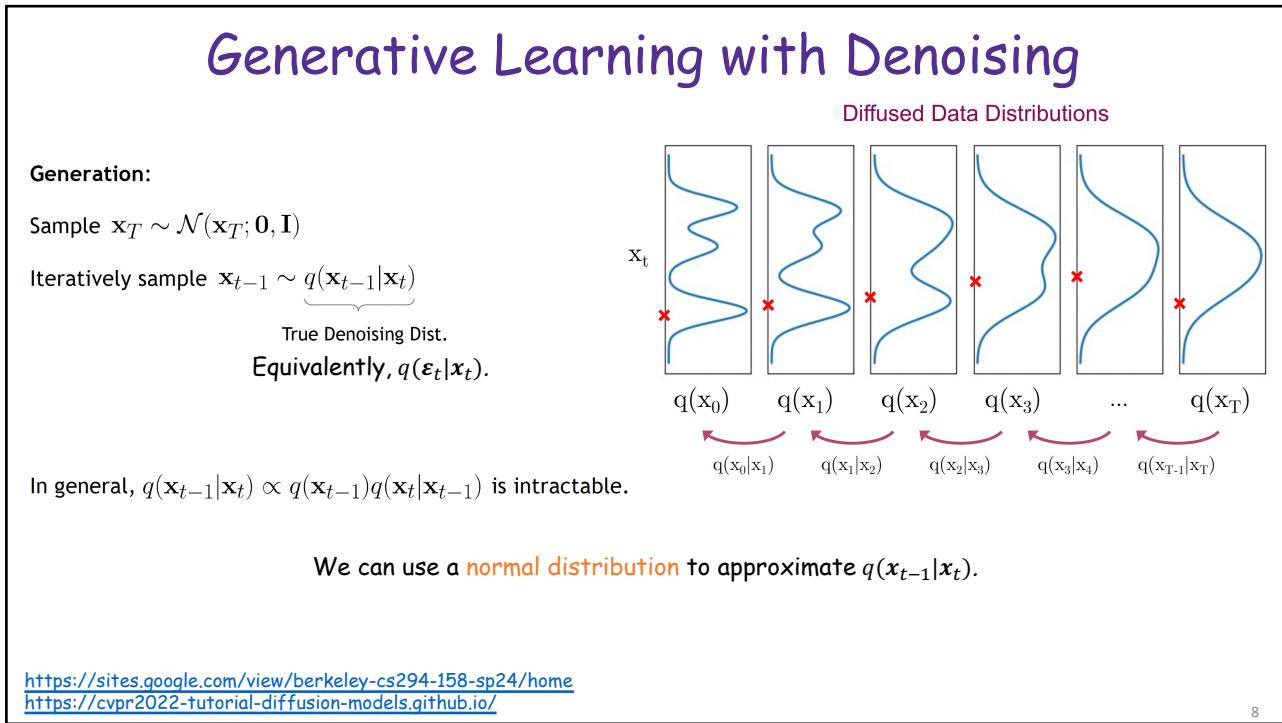
6

6



7

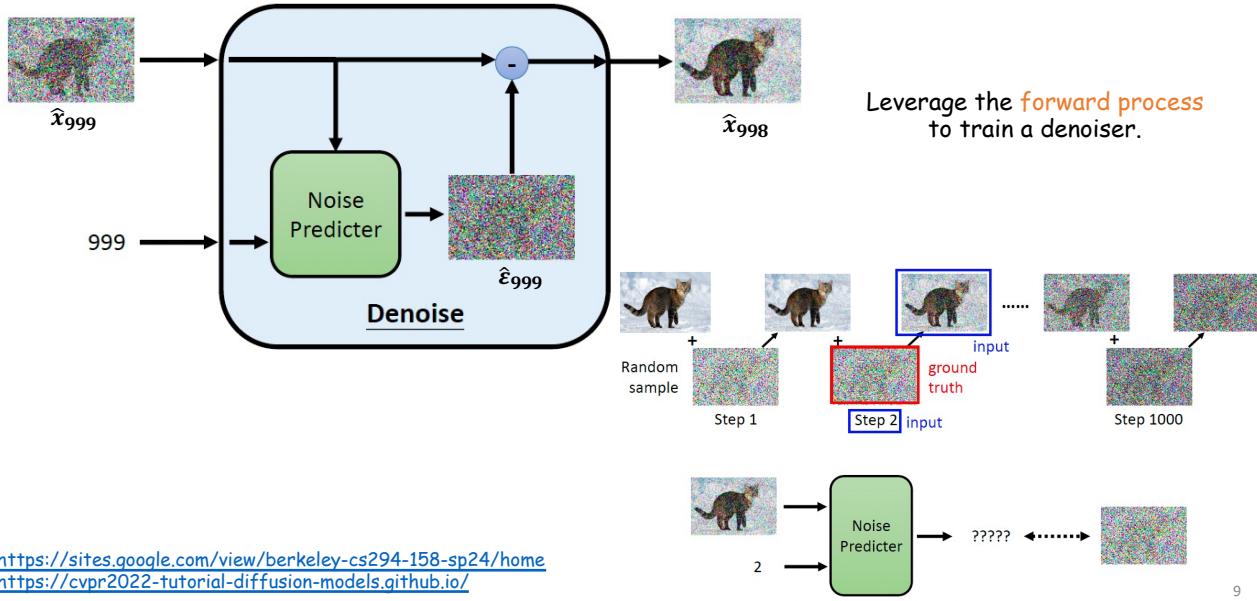
7



8

8

Training Denoisers



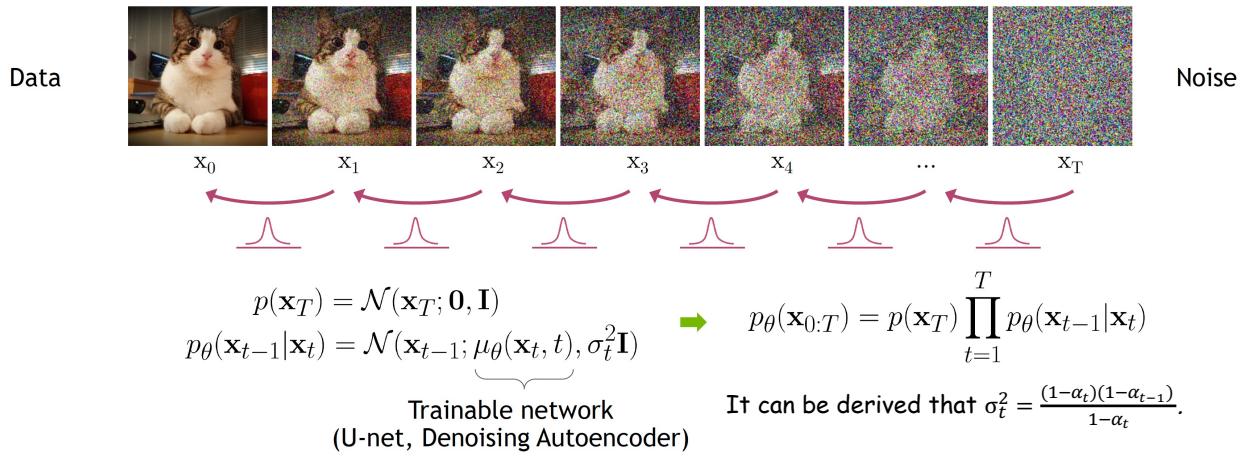
6

9

9

Reverse Denoising Process

Reverse denoising process (generative)



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

10

10

Evidence Lower Bound (ELBO)

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

- **Reconstruction term:** How good the neural nets $p_{\theta}(\mathbf{x}_0|\mathbf{x}_1)$ is to recover \mathbf{x}_0 from \mathbf{x}_1 when the sample is drawn from $q(\mathbf{x}_1|\mathbf{x}_0)$.
- **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 $p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ approximates the tractable ground-truth denoising transition step distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$.

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

- Train the noise $\widehat{\epsilon}_{\theta}(\mathbf{x}_t)$, which is a neural network that maps from \mathbf{x}_t to ϵ_0 , where $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon_0$

$$\text{ELBO}_{\theta} = - \sum_{t=1}^T \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2 \bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2} \|\widehat{\epsilon}_{\theta}(\mathbf{x}_t) - \epsilon_0\|^2 \right]. \quad \text{Then, sample } \mathbf{x}_{t-1} \text{ from } q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0).$$

<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:    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:   Take gradient descent step on
        $\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon, t)\|^2$ 
6: until converged

```

Algorithm 2 Sampling

```

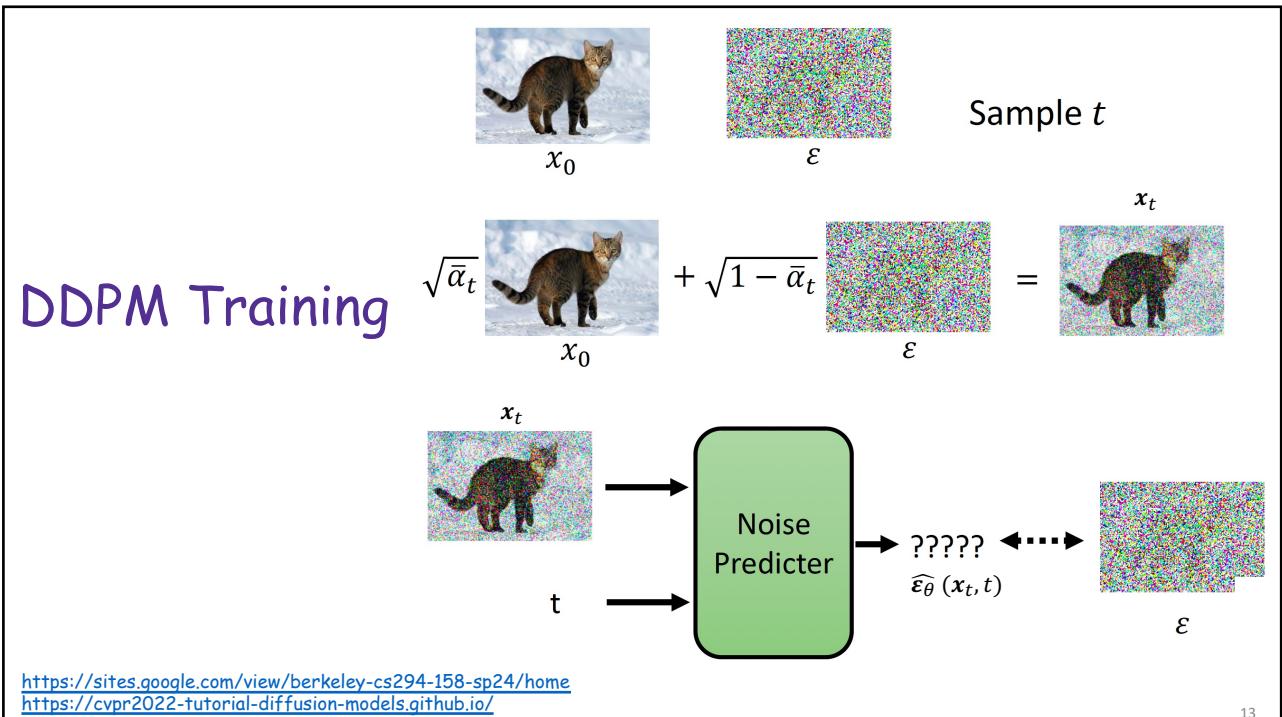
1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T, \dots, 1$  do
3:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  if  $t > 1$ , else  $\mathbf{z} = \mathbf{0}$ 
4:    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}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

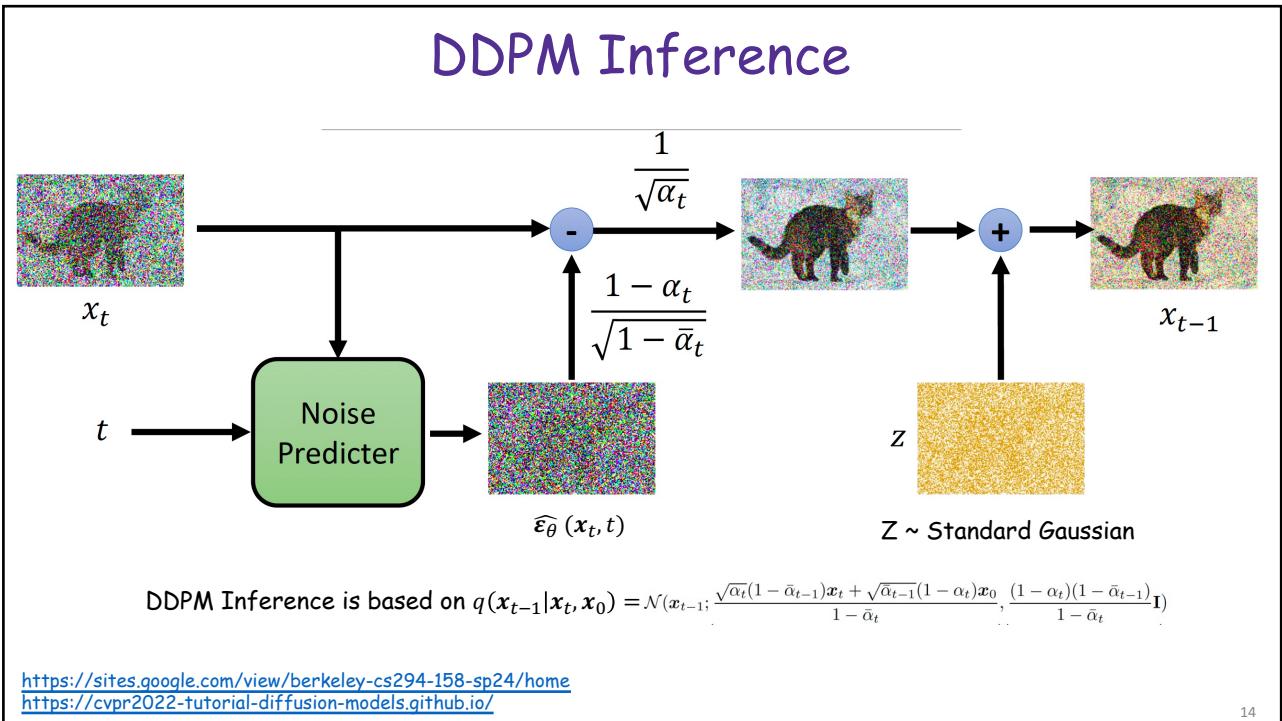
12

12



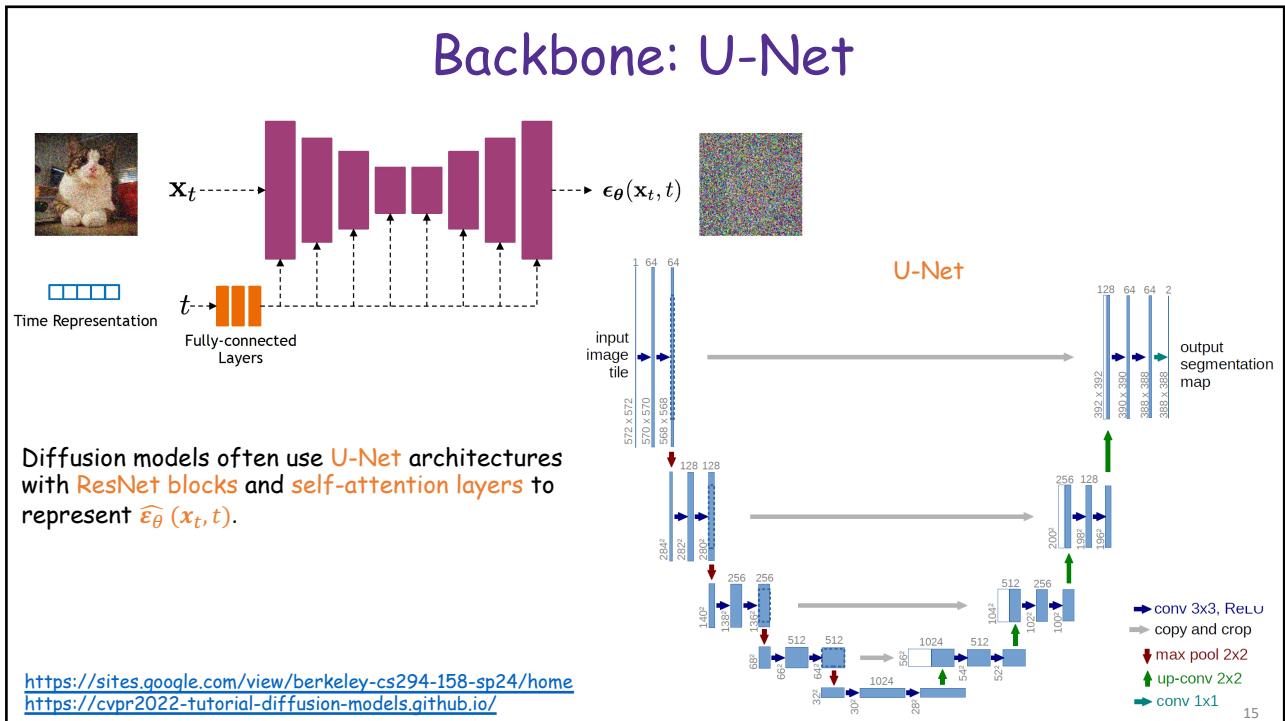
13

13



14

14



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 Σ)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	3.17	≤ 3.75 (3.72)

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

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

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].

17

<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**.

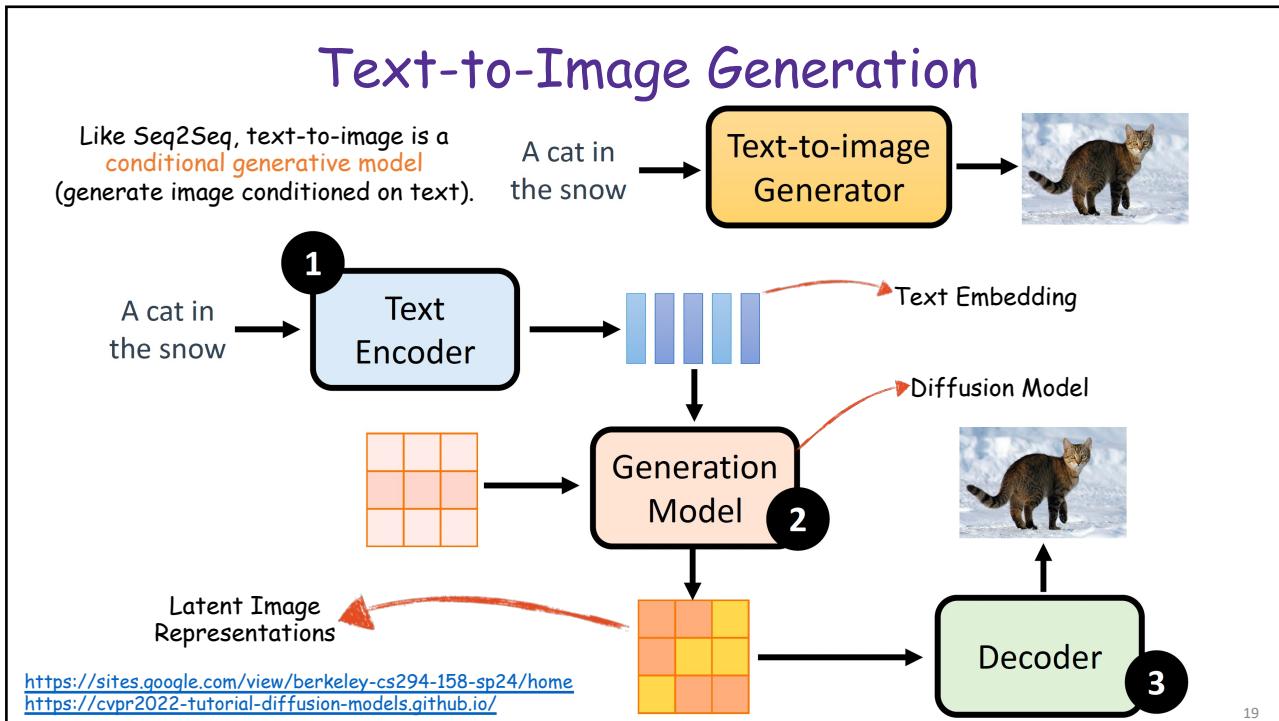
17

Agenda

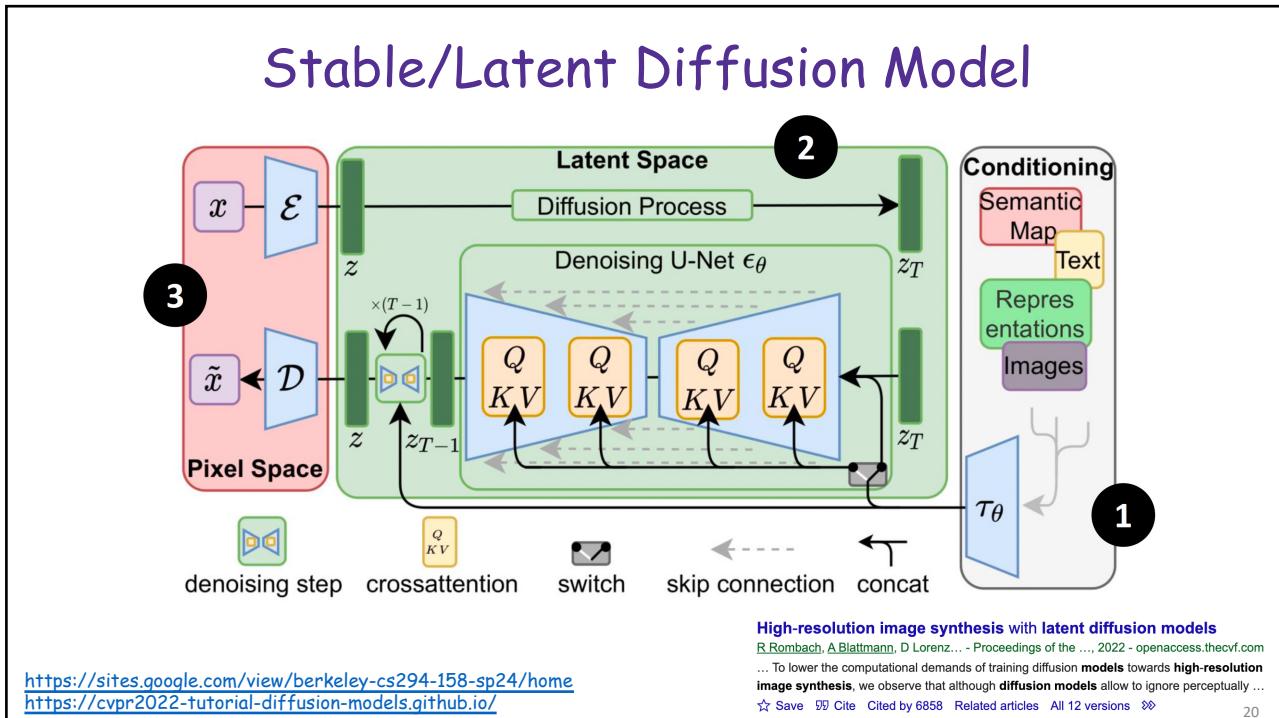
- Denoised Diffusion Probabilistic Models
- Latent Diffusion, CLIP, DALL-E. Diffusion Transformer
- Potential Applications of Diffusions in Biz/Econ Research

18

18



19



20

Stable Diffusion Results



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 ⚡ 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				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	3.08	0.65	0.46
<i>LDM-4</i> (ours, 500-s [†])				<i>LDM-4</i> (ours, 200-s)			
	5.11	0.72	0.49		4.98	0.73	0.50
LSUN-Churches 256 × 256							
Method	FID ↓	Prec. ↑	Recall ↑	Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN2 [42]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	0.66	0.51
ProjectedGAN [76]	1.59	0.61	0.44	ProjectedGAN [76]	1.52	0.61	0.34
<i>LDM-8</i> (ours, 200-s)				<i>LDM-4</i> (ours, 200-s)			
	4.02	0.64	0.52		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↑	Nparams
CogView [†] [17]	27.10	18.20	4B self-ranking, rejection rate 0.017
LAFTIE [†] [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
<i>LDM-KL-8</i>	23.31	20.03 ± 0.33	1.45B 250 DDIM steps
<i>LDM-KL-8-G*</i>	12.63	30.29 ± 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)

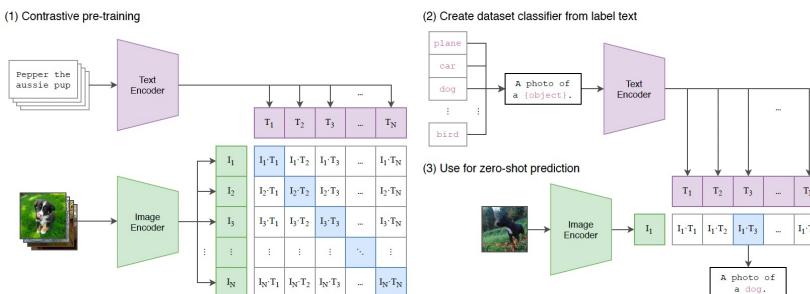
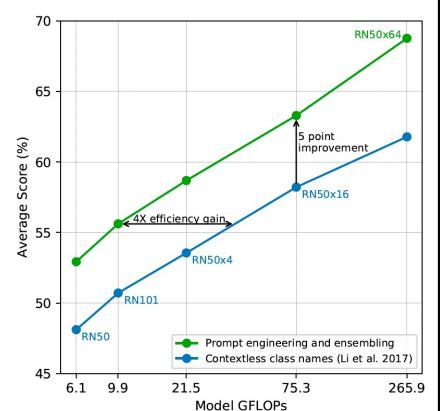


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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



Learning transferable visual models from natural language supervision
A Radford, JW Kim, C Hallacy... - ... machine learning, 2021 - proceedings.mlr.press

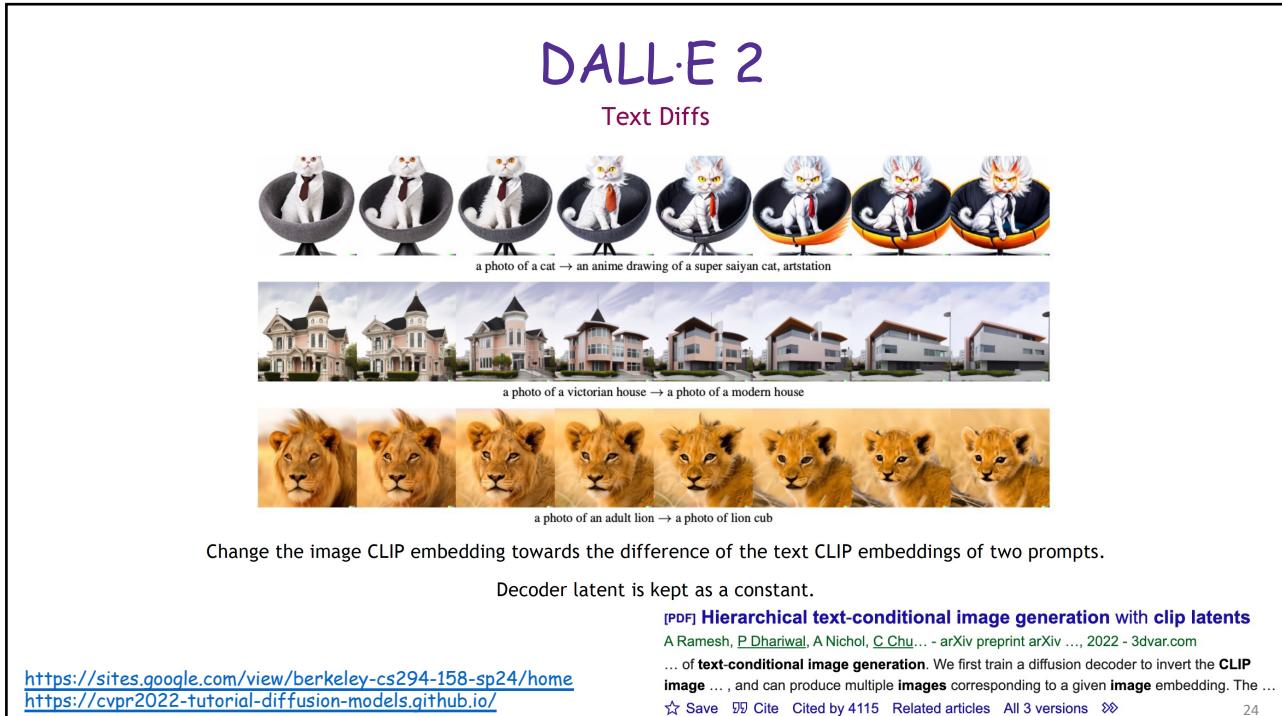
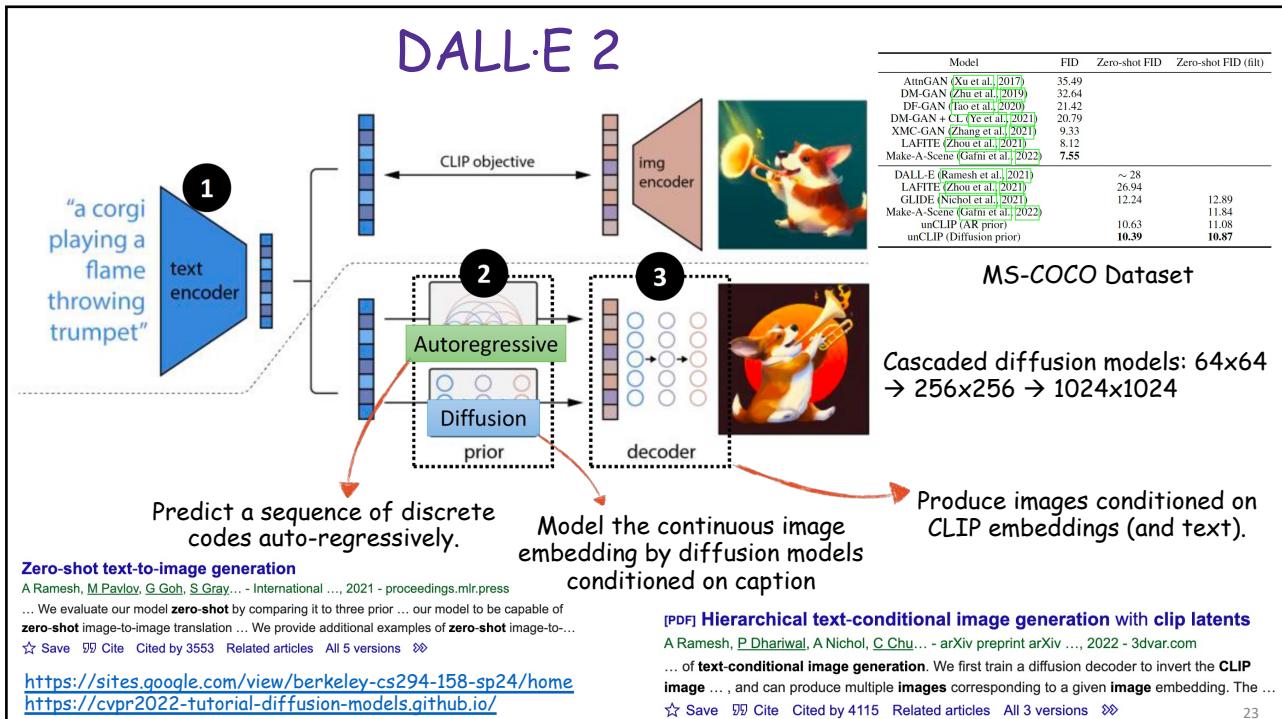
... We speculate this is due to natural language providing wider supervision for visual concepts involving verbs, compared to the noun-centric object supervision in ImageNet. ...
☆ 保存 ⚡ 引用 被引用次数：14959 相关文章 所有 19 个版本 ☺

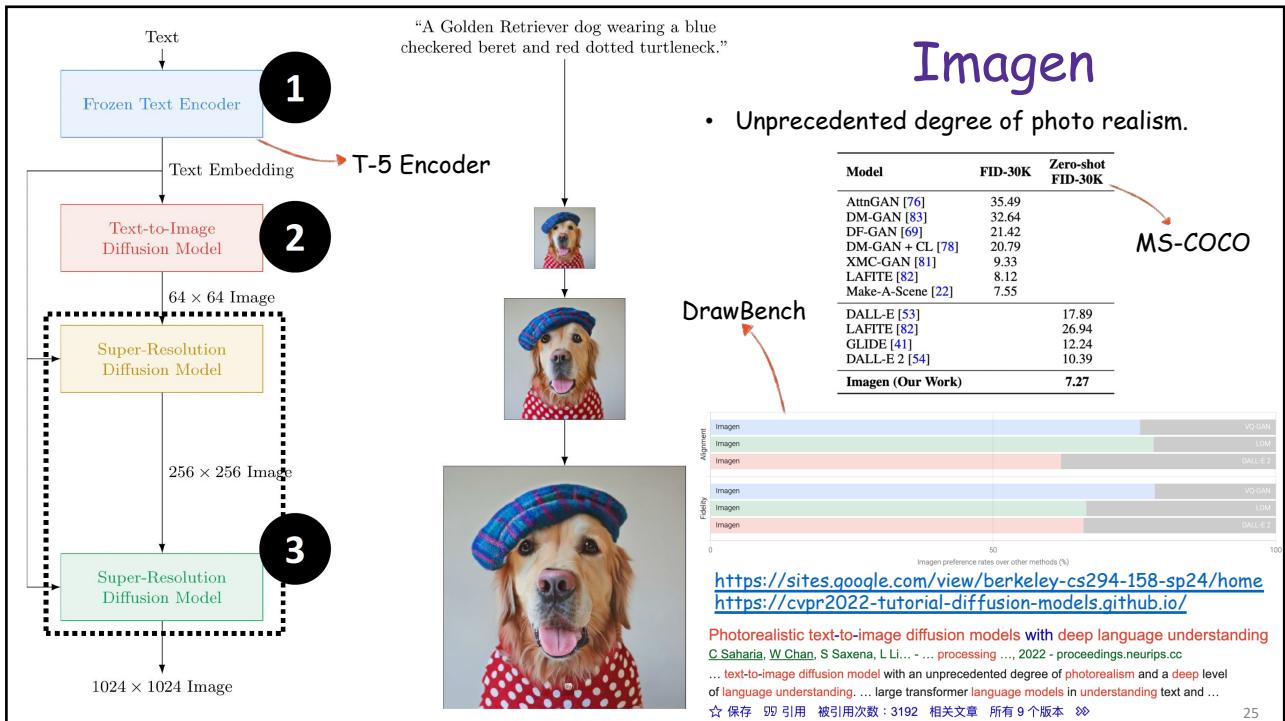
22

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

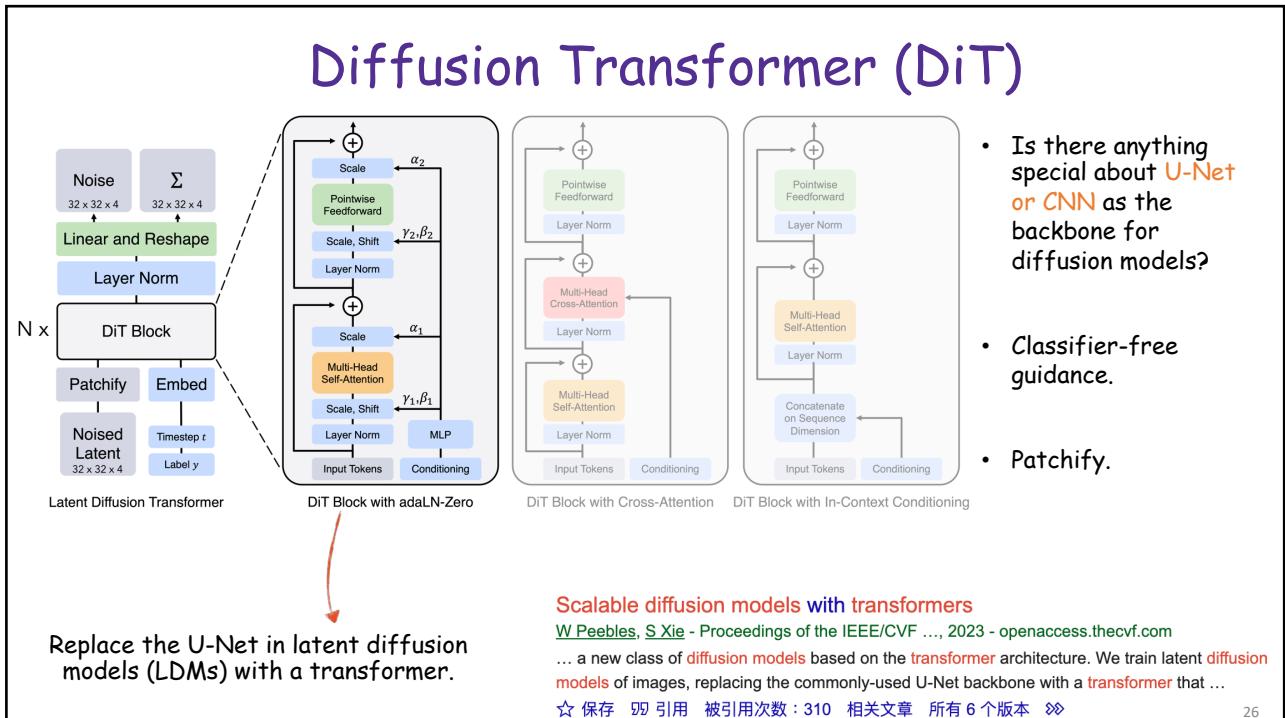
22

11



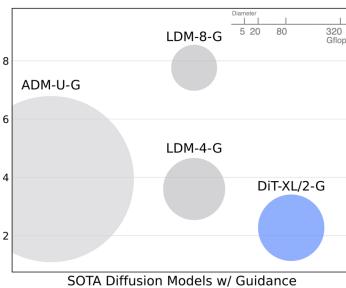
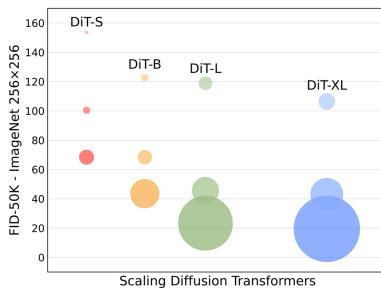
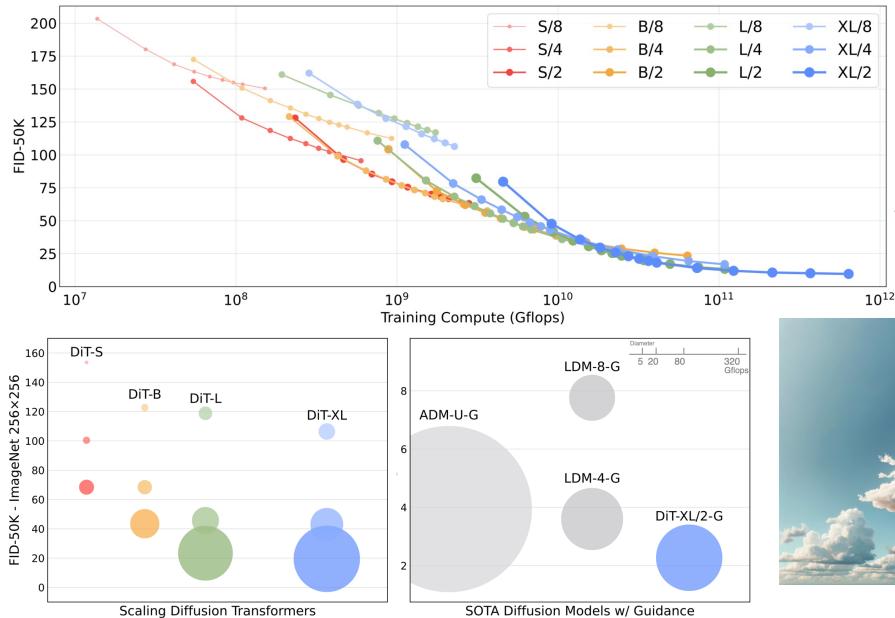


25



26

Diffusion Transformer (DiT) Performances



27

Agenda

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

28

28

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 Cite 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?

29

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 Related articles All 11 versions

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.

30

30

Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design
R Dew, A Ansari, O Toubla
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.



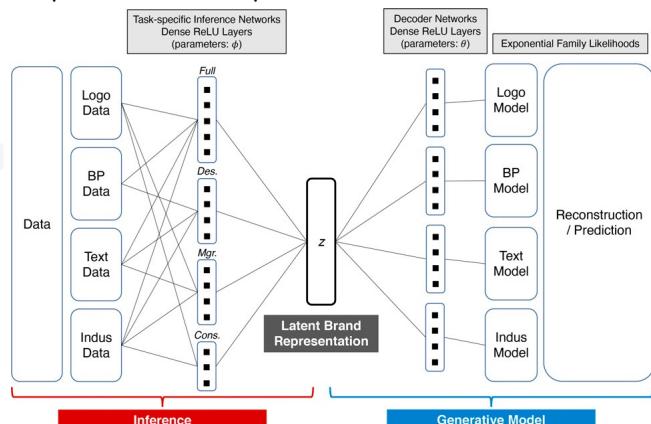
SHOW LESS ^

[☆ Save](#) [59 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 to learn the latent representation of logos, and to generate new logos and predict consumer preferences.



31

31

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.



SHOW LESS ^

[☆ Save](#) [59 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.

32

32