# Frontiers in Diffusion Model Technologies (1)

XIN GAO 2024.12.8

## **Content**

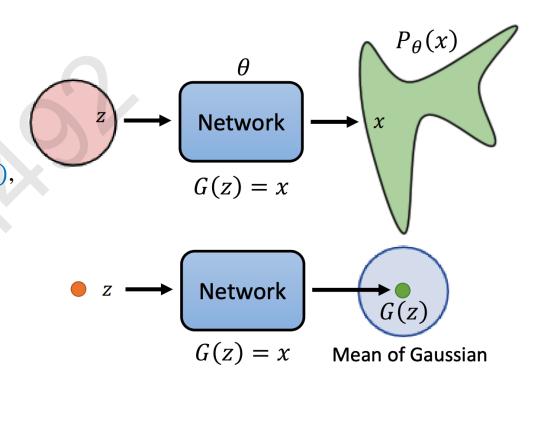
- Theoretical foundation :
  - DDPM
  - DDIM
  - SDE and ODE
  - Conditional Guidance
- Development Timeline
- Stable diffusion
  - Latent Diffusion
  - VQ-VAE
  - DiT
- Latest Methodology: IC-Light (ICLR 2025)

## VAE and ELBO

- A **VAE** models the distribution  $p_{data}(x)$  of the observed variable  $x \in \mathbb{R}^n$  by jointly learning a stochastic latent variable  $z \in \mathbb{R}^m$ .
- **Generation** is performed by sampling z from the prior p(z), then sampling x according to a probabilistic **decoder**.  $p_{\theta}(x|z)$  parametrized by  $\theta \in \Theta$ .
- How to update heta ? MLE  $p_{ heta}(x) = \int_{z}^{z} p(z) p_{ heta}(x \mid z) dz$
- Identity:  $q(z)\log p(x)dz$

Evidence 
$$= \int q(z) \log \frac{p_{\theta}(x\mid z)p(z)}{p(z\mid x)} \frac{q(z)}{q(z)} dz$$
For arbitrary 
$$= \int q(z) \log p_{\theta}(x\mid z) dz - KL[q(z)||p(z)] + KL[q(z)||p(z\mid x)]$$
of  $z$ 

Evidence Lower Bound (ELBO)



$$-KL[q(z)||p(z\mid x)]$$

KL divergence

### VAE and ELBO

Do a little math

- $\log p(x) \ge \text{ELBO}$  (KL divergence  $\ge 0$ ) Maximize ELBO => Increase  $\log p(x)$
- What is q(z)?

If 
$$q(z) = p(z|x)$$
,  $KL = 0$ ,  $\log p(x) = \text{ELBO}$  (EM Algorithm)  
Unfortunately, the true posterior  $p(z|x)$  is intractable,  $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$ 

• We use an encoder network to approximate the posterior

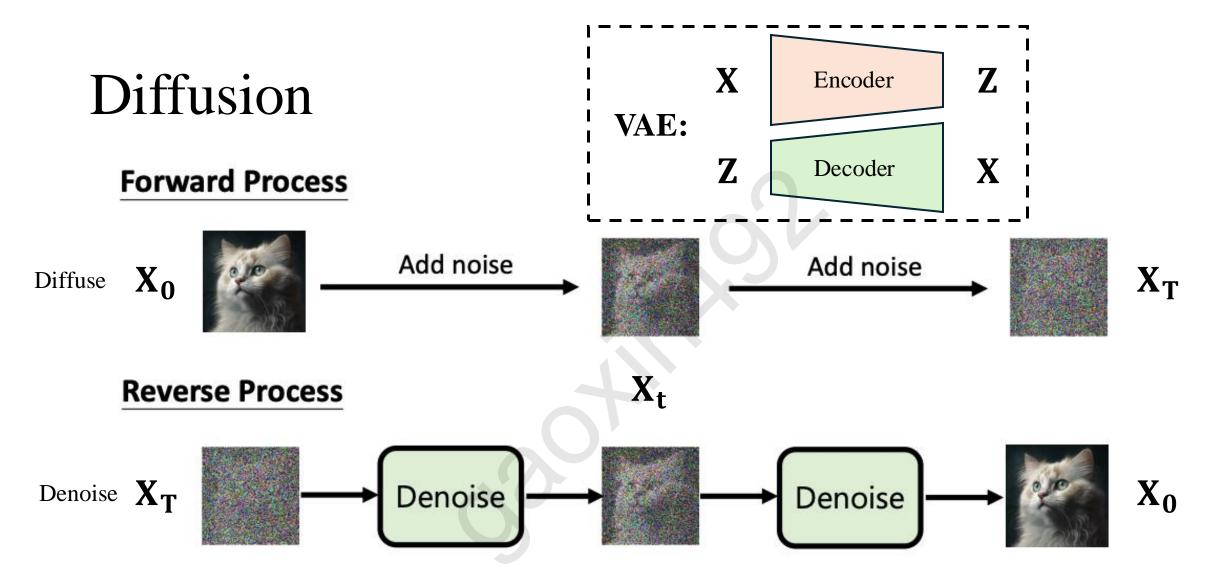
$$q_{\phi}(z|x) = \mathcal{N}(z; \mu(x), \Sigma(x))$$

By replacing q(z) with q(z|x), maximizing ELBO not only minimizes KL but also approximates MLE

$$\log p(x) = \int q_\phi(z|x) \log p_ heta(x\mid z) dz - KL[q_\phi(z|x)||p(z)] + KL[q_\phi(z|x)||p(z\mid x)]$$

• Objective:  $\mathbb{E}_{q_{\phi}(z|x)}[\log p_{ heta}(x\mid z)] - KL[q_{\phi}(z\mid x)\|p(z)]$ 

$$egin{align} ELBO &= \int q(z|x) \log rac{p(x,z)}{q(z|x)} dz \ &= \mathbb{E}_z[\log rac{p(x,z)}{q(z|x)}] \ &= \int q(z|x) \log rac{p(x\mid z)p(z)}{q(z|x)} dz \ &= \int q(z|x) \log p(x\mid z) dz - KL[q(z|x)||p(z)] \ &= \mathbb{E}_z[\log p(x\mid z)] - KL[q(z|x)||p(z)] \ \end{cases}$$



**Diffusion models create data from noise by inverting the forward paths of data towards noise** and have emerged as a powerful generative modeling technique for high-dimensional, perceptual data such as images and videos.

# DDPM Denoising Diffusion Probabilistic Model

- Original image  $x_0$
- Step-by-step decomposition, assuming multiple latent variables,  $p(x_{1:T}|x_0) := \prod_{t=1} p(x_t|x_{t-1})$ Markov chain  $x_0 \to x_1 \to x_2 \to ... \to x_T$
- Forward Process with decreasing sequence  $1 \ge \alpha_1 \ge \alpha_2 \ge ... \ge \alpha_T \ge 0$ ,  $\beta_t \coloneqq 1 \alpha_t$

$$x_t = \sqrt{lpha_t} x_{t-1} + \sqrt{(1-lpha_t)} arepsilon_t, \; arepsilon_t \sim \mathcal{N}(0,1), \; t = 1,...,T$$

Variable substitution / reparameterization trick  $p(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})$ 

Recursion (Noise  $\bar{\varepsilon}_t$ , linear combination of Gaussians still results in a Gaussian)

$$egin{aligned} x_t = \sqrt{ar{lpha}_t} x_0 + \sqrt{(1-ar{lpha}_t)} ar{arepsilon}_t, ext{ and } p(x_t|x_0) = \mathcal{N}(x_t; \sqrt{ar{lpha}_t} x_0, (1-ar{lpha}_t) \mathbf{I}) \end{aligned} \qquad ar{lpha}_t := \prod_{s=1}^t lpha_s$$

- When T steps are large enough  $\lim_{t \to +\infty} \bar{\alpha}_t = \prod_{s=1}^t \alpha_s = 0$   $p(x_T) \to \mathcal{N}(0,1)$
- How do we reconstruct the image step by step?

## DDPM Denoising Diffusion Probabilistic Model

• Bayes' Rule:  $p(x_{t-1}|x_t) = \frac{p(x_t|x_{t-1})p(x_{t-1})}{p(x_t)}$ 

But we do not know  $p(x_{t-1}), p(x_t)$ 

• We know conditional the distribution given  $x_0$ 

$$p(x_{t-1}|x_t,x_0) = rac{p(x_t|x_{t-1})p(x_{t-1}|x_0)}{p(x_t|x_0)}$$

 $p(x_{t-1}|x_t,x_0)=rac{p(x_t|x_{t-1})p(x_{t-1}|x_0)}{p(x_t|x_0)}$  Known Gaussian distributions

We can easily derive that  $p(x_{t-1}|x_t,x_0) = \mathcal{N}\left(x_{t-1}; \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \mathbf{I}\right)$ 

• But there's a gap. We can use  $x_t$  to predict/estimate  $x_0$ ,  $||x_0 - \mu_{\theta}(x_t)||^2$ 

$$p(x_{t-1}|x_t)pprox p(x_{t-1}|x_t,\hat{x}_0), \quad ext{where} \quad \hat{x}_0=\mu_{ heta}(x_t)$$

By making a small adjustment, due to  $x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{(1-\bar{\alpha}_t)}\bar{\varepsilon}_t)$ 

Predict the noise instead  $\mu_{\theta}(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{(1-\bar{\alpha}_t)}\epsilon_{\theta}(x_t,t))$ 

# DDPM Denoising Diffusion Probabilistic Model

$$\log p(x) \ge \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{0:T})}{q(x_{1:T}|x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) \prod_{t=1}^{T} p_{\theta}(x_{t-1}|x_{t})}{\prod_{t=1}^{T} q(x_{t}|x_{t-1})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1}) \prod_{t=2}^{T} p_{\theta}(x_{t-1}|x_{t})}{q(x_{1}|x_{0}) \prod_{t=2}^{T} q(x_{t}|x_{t-1})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1}) \prod_{t=2}^{T} p_{\theta}(x_{t-1}|x_{t})}{q(x_{1}|x_{0}) \prod_{t=2}^{T} q(x_{t}|x_{t-1},x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{1}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t}|x_{t-1},x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{1}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})q(x_{t}|x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{1}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})q(x_{t}|x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{1}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})q(x_{t}|x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{T}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})q(x_{t}|x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{T}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{T}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{T}|x_{0})} + \log \frac{p(x_{T}|x_{0})}{q(x_{T}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{t-1}|x_{t},x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{T}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{T}|x_{0})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_{0})} \left[ \log \frac{p(x_{T}) p_{\theta}(x_{0}|x_{1})}{q(x_{T}|x_{0})} + \log \prod_{t=2}^{T} \frac{p_{\theta}(x_{t-1}|x_{t})}{q(x_{T}|x_{0})} \right]$$

- $= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[ \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[ \log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right] + \sum_{t=0}^T \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[ \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right]$ (56)
- $= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[ \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[ \log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right] + \sum_{t=0}^{T} \mathbb{E}_{q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \left[ \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} \right] \right]$ (57)

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{1}|\boldsymbol{x}_{0})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0}|\boldsymbol{x}_{1})\right]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q(\boldsymbol{x}_{T}|\boldsymbol{x}_{0}) \parallel p(\boldsymbol{x}_{T}))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t}|\boldsymbol{x}_{0})}\left[D_{\text{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_{t},\boldsymbol{x}_{0}) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_{t}))\right]}_{\text{denoising matching term}}$$
(58)

• ELBO

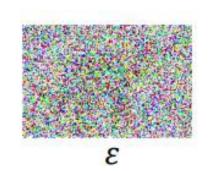
(55)

- Perspective from Latent Variable Model (like VAE)
- 知道就行,这种就是从隐变量模型 出发, 推导 ELBO 得到 loss, 最终 loss带入分布后化简得到相同的结 果。但中间有一步推导比较 trick, 不如前两页的好理解
- 其实 Diffusion 就是一个中间隐变量 是层级建模的VAE (Hierarchical VAE) + 将 encode 过程确定为了扩 散过程 instead of learnable encoder

## **Training**

 $\bar{\alpha}_1, \bar{\alpha}_2, \dots \bar{\alpha}_T$ 

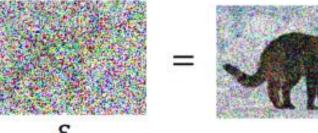




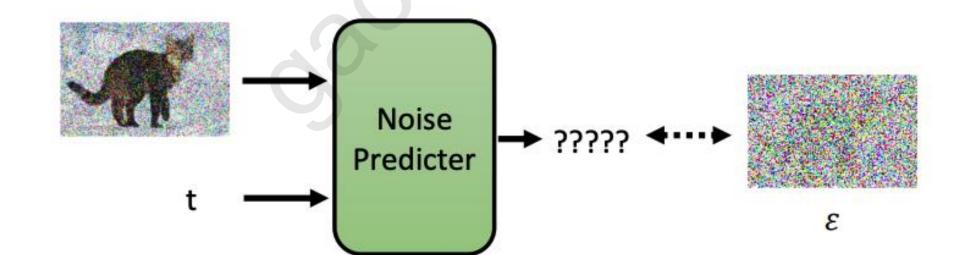
Sample t

$$\sqrt{\bar{\alpha}_t}$$

$$+\sqrt{1-\bar{\alpha}_t}$$







```
# construct DDPM noise schedule
b_t = (beta2 - beta1) * torch.linspace(0, 1, timesteps + 1, device=device) + beta1
a_t = 1 - b_t
ab_t = torch.cumsum(a_t.log(), dim=0).exp()
ab_t[0] = 1
```

https://github.com/Ryota-Kawamura/How-Diffusion-Models-Work/tree/main

```
# helper function: perturbs an image to a specified noise level

def perturb_input(x, t, noise):
    return ab_t.sqrt()[t, None, None, None] * x + (1 - ab_t[t, None, None, None]).sqrt() * noise
```

```
set into train mode
nn_model.train()
for ep in range(n_epoch):
    print(f'epoch {ep}')
    # linearly decay learning rate
    optim.param_groups[0]['lr'] = lrate*(1-ep/n_epoch)
    pbar = tqdm(dataloader, mininterval=2 )
    for x, _ in pbar: # x: images
        optim.zero_grad()
        x = x.to(device)
        # perturb data
        noise = torch.randn_like(x)
        t = torch.randint(1, timesteps + 1, (x.shape[0],)).to(device)
        x_pert = perturb_input(x, t, noise)
        # use network to recover noise
        pred_noise = nn_model(x_pert, t / timesteps)
        # loss is mean squared error between the predicted and true noise
        loss = F.mse_loss(pred_noise, noise, reduction='sum') / x.shape[0]
        print(f'loss: {loss.item():.4f}', end='\r')
        loss.backward()
        optim.step()
```



 $x_0$ : clean image



€: noise

### Algorithm 1 Training

- 1: repeat
- 2:  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  --- sample clean image
- 3:  $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4:  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  ---- sample a noise
- 5: Take gradient descent step on

6: **until** converged Noisy image 
$$\overline{\alpha}_1, \overline{\alpha}_2, ... \overline{\alpha}_T$$
 smaller

Target Noise Noise predictor

# DATASET

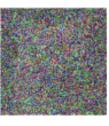
# MODEL

INP Noise Amount		OUTPUT / LABEL Noise sample
3		
14		
7		
42		
2	802	
21		

Noise Predictor

### Inference

### **Algorithm 2** Sampling



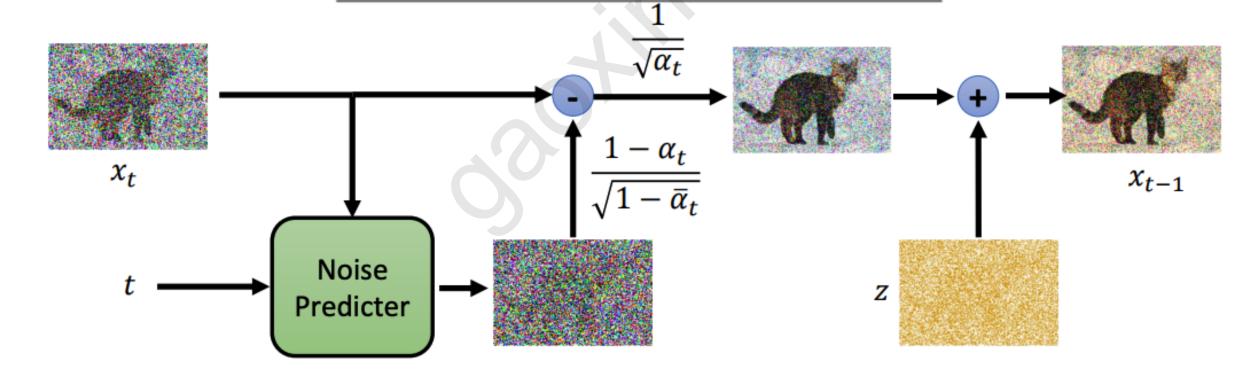
 $x_T$ 

- 1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 2: **for** t = T, ..., 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**  $\mathbf{x}_0$

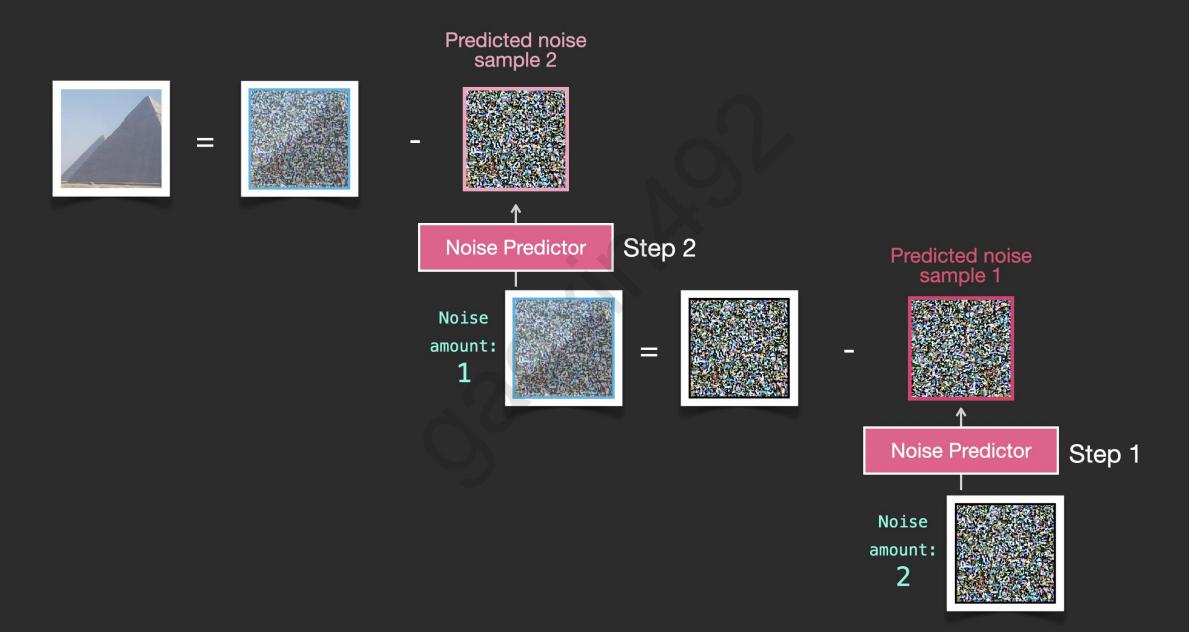
sample a noise?!

 $\bar{\alpha}_1, \bar{\alpha}_2, \dots \bar{\alpha}_T$ 

 $\alpha_1, \alpha_2, \dots \alpha_T$ 



### Image Generation by Reverse Diffusion (Denoising)

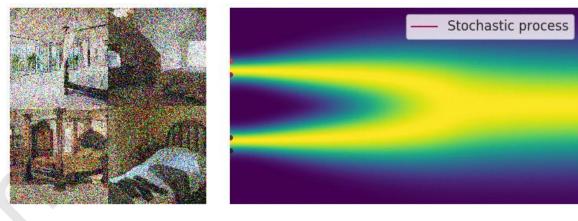


# Stochasticity

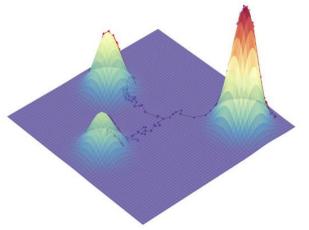
### Think again about the stochasticity

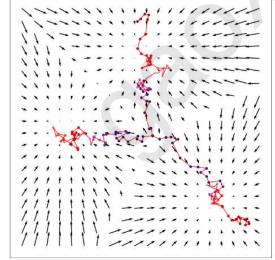
$$x_{t-1} = rac{1}{\sqrt{lpha_t}} \left( x_t - rac{1-lpha_t}{\sqrt{1-arlpha_t}} oldsymbol{\epsilon}_ heta(x_t,t) 
ight) + \sigma_t oldsymbol{arepsilon}$$

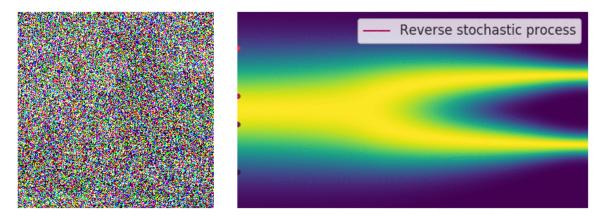
# Sampling from a distribution!



Perturbing data to noise with a continuous-time stochastic process.







Generate data from noise by reversing the perturbation procedure.

## DDIM Denoising Diffusion Implicit Model

$$egin{aligned} p(oldsymbol{x}_t|oldsymbol{x}_{t-1}) & \stackrel{ ext{ iny property}}{\longrightarrow} p(oldsymbol{x}_t|oldsymbol{x}_0) & \stackrel{ ext{ iny property}}{\longrightarrow} p(oldsymbol{x}_{t-1}|oldsymbol{x}_t,oldsymbol{x}_0) & \stackrel{ ext{ iny property}}{\longrightarrow} p(oldsymbol{x}_{t-1}|oldsymbol{x}_t) & \chi_0 & \\ & & \text{Forward/Diffuse} & \text{Reverse/Denoise} & & \\ \end{aligned}$$

• **Training**: The loss only relies on  $p(x_t|x_0)$ 

$$rac{1-ar{lpha}_t}{ar{lpha}_t}||oldsymbol{arepsilon}-oldsymbol{\epsilon}_{ heta}(x_t,t)||^2=rac{1-ar{lpha}_t}{ar{lpha}_t}||oldsymbol{arepsilon}-oldsymbol{\epsilon}_{ heta}(\sqrt{ar{lpha}}_toldsymbol{x}_0+\sqrt{1-ar{lpha}_t}oldsymbol{arepsilon},t)||^2$$

• Sampling: Each step sampling only relies on  $p(x_{t-1}|x_t)$ 

Maybe we do not need to set  $p(x_t|x_{t-1})$  and assume Markov chain process?

$$p(x_{t-1}|x_t,x_0) = rac{p(x_t|x_{t-1})p(x_{t-1}|x_0)}{p(x_t|x_0)} \;\; (*) \;\;\;\;\;\;\;\; \int p(x_{t-1}|x_t,x_0)p(x_t|x_0)dx_t = p(x_{t-1}|x_0) \;\;\; (**)$$

• Actually we have more distributions  $p(x_{t-1}|x_t,x_0)$  to satisfy Eq. (\*\*)
Undetermined Coefficients  $\kappa_t$ ,  $\lambda_t$ ,  $\sigma_t$ ,  $p(x_{t-1}|x_t,x_0) = \mathcal{N}(x_{t-1};\kappa_t x_t + \lambda_t x_0,\sigma_t^2 \mathbf{I})$ 

# **DDIM** Denoising Diffusion Implicit Model

$$egin{aligned} x_0 & p(oldsymbol{x}_t|oldsymbol{x}_{t-1}) & ext{ ilde{\#}} p(oldsymbol{x}_t|oldsymbol{x}_0) & ext{ ilde{\#}} p(oldsymbol{x}_{t-1}|oldsymbol{x}_t,oldsymbol{x}_0) & ext{ ilde{\text{ iny figure}}} p(oldsymbol{x}_{t-1}|oldsymbol{x}_t,oldsymbol{x}_0) & ext{ iny figure} & ext{ i$$

$$egin{aligned} x_t = \sqrt{ar{lpha}_t} x_0 + \sqrt{(1-ar{lpha}_t)} ar{arepsilon}_t, ext{ and } p(x_t|x_0) = \mathcal{N}(x_t; \sqrt{ar{lpha}_t} x_0, (1-ar{lpha}_t) \mathbf{I}) \end{aligned}$$

$$\int p(x_{t-1}|x_t,x_0) p(x_t|x_0) dx_t = p(x_{t-1}|x_0) \quad (**) \qquad p(x_{t-1}|x_t,x_0) = \mathcal{N}(x_{t-1}; rac{\kappa_t}{\kappa_t} x_t + rac{\lambda_t}{\lambda_t} x_0, rac{\sigma_t^2}{\sigma_t^2} \mathbf{I})$$

定好的超参数,是已知的

• **DDPM**: 
$$\sigma_t^2 = rac{1-ar{lpha}_{t-1}}{1-ar{lpha}_t}eta_t$$

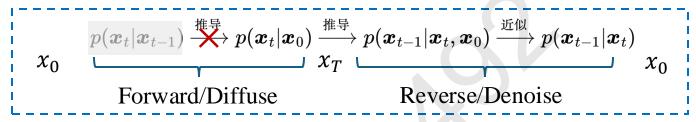
• **DDIM**: 
$$\sigma_t^2 = 0$$
 Implicit 隐式的概率模型,确定性采样过程,不带随机性  $\frac{1}{1}$  Remark: 在给定  $p(x_{t-1}|x_t,x_0)$  后,我们还

$$p(x_{t-1}|x_t,x_0) = rac{p(x_t|x_{t-1})p(x_{t-1}|x_0)}{p(x_t|x_0)}$$

Larger covariance:  $\sigma_t^2 = \beta_t$ 

## DDIM Denoising Diffusion Implicit Model

### Accelerated Generation Process



$$egin{aligned} x_t = \sqrt{ar{lpha}_t} x_0 + \sqrt{(1-ar{lpha}_t)} ar{arepsilon}_t, ext{ and } p(x_t|x_0) = \mathcal{N}(x_t; \sqrt{ar{lpha}_t} x_0, (1-ar{lpha}_t) \mathbf{I}) \end{aligned}$$

$$\text{Given} \ \ \boldsymbol{\sigma}_t: \ \ \boldsymbol{\kappa}_t = \frac{\sqrt{\bar{\beta}_{t-1} - \sigma_t^2}}{\sqrt{\bar{\beta}_t}}, \quad \boldsymbol{\lambda}_t = \sqrt{\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_t}\sqrt{\bar{\beta}_{t-1} - \sigma_t^2}}{\sqrt{\bar{\beta}_t}} \\ p(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \frac{\boldsymbol{\kappa}_t x_t + \boldsymbol{\lambda}_t x_0, \frac{\sigma_t^2}{\sigma_t^2} \mathbf{I})}{\sqrt{\bar{\beta}_t}}$$

Suppose that an increasing subsequence of [1, ..., T]:  $[\tau_1, ..., \tau_S]$ 

$$egin{aligned} p(oldsymbol{x}_{ au_t}|oldsymbol{x}_0) & frac{ f f f}}}{ frac{ frac{ frac{ frac{ frac{ frac{ frac{ frac{1}}{ frac{ f{x}}}}{ frac{ frac{ frac{ frac{ frac{ f{x}}}{ frac{ frac{ frac{ f{x}}}{ frac{ frac{ frac{ frac{ frac{1}}{ frac{ frac{ f{ta}}}}{ frac{ frac{ f{ta}}}{ frac{ frac{ f{x}}}{ frac{ frac{ f{ f f f{x}}}}{ frac{ frac{ f{ta}}}}{ frac{ f{x}}}}{ frac{ frac{ f{x}}}{ frac{ f{x}}}}{ frac{ frac{ f{x}}}{ frac{ f{x}}}{ frac{ f{x}}}}{ frac{ f{x}}}{ fita}}}}}}}}{ frac{ frac{ frac{ f{f}}}{ frac{ f{x}}{ frac{ f{x}}{\frac{ f{x}}}{ frac{ f{x}}}{ frac{1}}{ frac{ f{x}}}{ frac{ f{x}}{ fita}}}}{ frac{ frac{ f{f}}}{ frac{ f{x}}{ fr$$

It is allowed to skip steps! Original 1000 steps, 10 steps per jump => 100 steps, 20 steps per jump => 50 steps

## SDE

Forward process in DDPM:  $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{(1-\alpha_t)\varepsilon_t}, \ \varepsilon_t \sim \mathcal{N}(0,1), \ t=1,...,T$ 

连续化 一般化 
$$x_{t+\Delta t}-x_t=m{f}_t(x_t)\Delta t+g_t\sqrt{\Delta t}m{arepsilon}, \quad m{arepsilon}\sim \mathcal{N}(0,\mathbf{I}), \quad \Delta t o 0$$

=> **SDE**:

$$dx = f_t(x)dt + g_t dw$$

 $dx = f_t(x)dt + g_t dw$  w: Wiener process or 布朗运动,是一个随机过程, 具有独立增量和连续轨迹,增量  $dw \sim \mathcal{N}(0, dt)$ 

- Drift coefficient  $f_t(x)dt$ : 系统的确定性变化
- Diffusion coefficient  $g_t dw$ :由随机扰动引起的不确定变化
- 概率分布形式  $p(x_{t+\Delta t}|x_t) = \mathcal{N}\left(x_{t+\Delta t}; x_t + \boldsymbol{f}_t(x_t)\Delta t, g_t^2 \Delta t \mathbf{I}\right) \propto \exp\left(-\frac{\|x_{t+\Delta t} x_t \boldsymbol{f}_t(x_t)\Delta t\|^2}{2\sigma^2 \Delta t}\right)$
- 逆向过程推导

$$egin{aligned} p(x_t|x_{t+\Delta t}) &= rac{p(x_{t+\Delta t}|x_t)p(x_t)}{p(x_{t+\Delta t})} = p(x_{t+\Delta t}|x_t) \exp(\log p(x_t) - \log p(x_{t+\Delta t})) \ &\propto \exp\left(-rac{||x_{t+\Delta t}-x_t-oldsymbol{f}_t(x_t)\Delta t||^2}{2g_t^2\Delta t} + \log p(x_t) - rac{\log p(x_{t+\Delta t})}{\log p(x_t)}
ight) \end{aligned}$$

 $\Delta t$  足够小,Taylor expansion:  $\log p(x_{t+\Delta t}) pprox \log p(x_t) + (m{x}_{t+\Delta t} - x_t) \cdot 
abla_{x_t} \log p(x_t) + \Delta t \frac{\partial}{\partial t} \log p(x_t)$ 

## SDE

**DDPM** 

正向  $\mathbf{SDE} \ | \ dx = f_t(x)dt + g_t dw$ 

$$x_t = \sqrt{lpha_t} x_{t-1} + \sqrt{(1-lpha_t)} arepsilon_t$$

$$egin{aligned} p(x_t|x_{t+\Delta t}) &\propto \exp\left(-rac{\|x_{t+\Delta t} - x_t - oldsymbol{f}_t(x_t)\Delta t\|^2}{2g_t^2\Delta t} + \log p(x_t) - \log p(x_{t+\Delta t})
ight) \ \log p(x_{t+\Delta t}) &pprox \log p(x_t) + (oldsymbol{x}_{t+\Delta t} - x_t) \cdot 
abla_{x_t} \log p(x_t) + \Delta t rac{\partial}{\partial t} \log p(x_t) \ p(x_t|x_{t+\Delta t}) &pprox \exp\left(-rac{\|x_{t+\Delta t} - x_t - oldsymbol{f}_t(x_t) - g_t^2 
abla_{x_t} \log p(x_t) ig] \Delta t\|^2}{2g_t^2\Delta t} + \mathcal{O}(\Delta t)
ight), \quad \Delta t o 0, \mathcal{O}(\Delta t) o 0 \ p(x_t|x_{t+\Delta t}) &pprox \exp\left(-rac{\|x_{t+\Delta t} - x_t - oldsymbol{f}_t(x_t) - g_t^2 
abla_{x_t} \log p(x_t) ig] \Delta t\|^2}{2g_t^2\Delta t}
ight) \ &pprox \exp\left(-rac{\|oldsymbol{x}_{t+\Delta t} - oldsymbol{x}_{t+\Delta t} - oldsymbol{f}_t(x_{t+\Delta t}) - g_{t+\Delta t}^2 
abla_{x_t+\Delta t} \log p(x_{t+\Delta t}) ig] \Delta t\|^2}{2g_{t+\Delta t}^2\Delta t}
ight) \end{aligned}$$

$$egin{aligned} x_t - x_{t+\Delta t} &= -\left[oldsymbol{f}_{t+\Delta t}(x_{t+\Delta t}) - g_{t+\Delta t}^2 
abla_{x_{t+\Delta t}} \log p(x_{t+\Delta t})
ight] \Delta t + g_{t+\Delta t} \sqrt{\Delta t} oldsymbol{arepsilon} \end{aligned}$$

逆向 SDE 
$$\Delta t o 0, \quad dx = [oldsymbol{f}_t(x) - g_t^2 
abla_x \log p_t(x)] dt + g_t doldsymbol{w}$$

$$\textbf{Loss:} \quad \mathbb{E}_{x_0, x_t \sim p(x_t \mid x_0) \tilde{p}(x_0)} \left[ \left\| \boldsymbol{s}_{\theta}(x_t, t) - \nabla_{x_t} \log p(x_t \mid x_0) \right\|^2 \right]$$

Loss 的推导本次省略

$$x_{t-1} = rac{1}{\sqrt{lpha_t}} \left( x_t - rac{1-lpha_t}{\sqrt{1-arlpha_t}} rac{oldsymbol{\epsilon}_ heta(x_t,t)}{oldsymbol{\epsilon}_ heta(x_t,t)} 
ight) + \sigma_t oldsymbol{arepsilon}$$

$$\mathbb{E}_{x_0,oldsymbol{arepsilon}}\left[rac{eta_t^2}{2\sigma_t^2lpha_t(1-ar{lpha}_t)}||oldsymbol{arepsilon}-oldsymbol{arepsilon}_ heta(\sqrt{ar{lpha}_t}x_0+\sqrt{1-ar{lpha}_t}oldsymbol{arepsilon},t)||^2
ight]$$

## SDE and ODE 大一统

$$dx = f_t(x)dt + g_t dw$$
 (#) Fokker-Planck 方程 描述边际分布的 PDE  $\frac{\partial}{\partial t} p_t(x) = -\nabla_x \cdot [\boldsymbol{f}_t(x)p_t(x)] + \frac{1}{2}g_t^2 \nabla_x \cdot \nabla_x p_t(x)$ 

对 FP 方程做等式变换,注意以下式子对  $\forall \sigma_t$  都成立:

$$egin{aligned} rac{\partial}{\partial t} p_t(x) &= -
abla_x \cdot \left[ oldsymbol{f}_t(x) p_t(x) - rac{1}{2} (g_t^2 - \sigma_t^2) 
abla_x p_t(x) 
ight] + rac{1}{2} \sigma_t^2 
abla_x \cdot 
abla_x p_t(x) \ &= -
abla_x \cdot \left[ \left( oldsymbol{f}_t(x) - rac{1}{2} (g_t^2 - \sigma_t^2) 
abla_x \log p_t(x) 
ight) p_t(x) 
ight] + rac{1}{2} \sigma_t^2 
abla_x \cdot 
abla_x p_t(x) \end{aligned}$$

我们发现这个 FP 方程也是以下 SDE 的 FP 方程:

$$dx = \left(oldsymbol{f}_t(x) - rac{1}{2}(g_t^2 - \sigma_t^2)
abla_x \log p_t(x)
ight)dt + \sigma_t doldsymbol{w} \quad ext{(##)}$$

也就是说式 (#) 和式 (##) 对应的marginal distribution  $p_t(x)$  完全相同即存在不同方差的前向过程,产生的 marginal distribution 完全相同

同样的,我们可以写出 (##) 的反向SDE: 
$$dx = \left(\boldsymbol{f}_t(x) - \frac{1}{2}(g_t^2 + \sigma_t^2)\nabla_x \log p_t(x)\right)dt + \sigma_t d\boldsymbol{w}$$

## SDE and ODE 大一统

$$dx = \left(oldsymbol{f}_t(x) - rac{1}{2}(g_t^2 - \sigma_t^2)
abla_x \log p_t(x)
ight)dt + \sigma_t doldsymbol{w} \quad ext{(##)}$$

$$dx = \left(oldsymbol{f}_t(x) - rac{1}{2}(g_t^2 + \sigma_t^2)
abla_x \log p_t(x)
ight)dt + \sigma_t doldsymbol{w}$$

### What if $\sigma_t = 0$ ?

### **Probability flow ODE**

$$dx = \left(oldsymbol{f}_t(x) - rac{1}{2}g_t^2
abla_x\log p_t(x)
ight)dt$$
 Deterministic transform

- Deterministic representation
- ODE Accelerated Solver Algorithm

**Remark:** The forward process and reverse process of ODE are exactly the same

## Score Function

Connecting gradient with the predicted noise:  $q(x_{t-1}|x_t) \approx \mathcal{N}(x_{t-1}; \frac{1}{\sqrt{\alpha_t}} \boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \sigma_t \mathbf{I})$ 

In this case, we apply it to predict the true posterior mean of  $x_t$  given its samples. From Equation 70, we know that:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

Then, by Tweedie's Formula, we have:

$$\mathbb{E}\left[\boldsymbol{\mu}_{x_t}|\boldsymbol{x}_t\right] = \boldsymbol{x}_t + (1 - \bar{\alpha}_t)\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t) \tag{131}$$

$$\sqrt{\bar{\alpha}_t} \boldsymbol{x}_0 = \boldsymbol{x}_t + (1 - \bar{\alpha}_t) \nabla \log p(\boldsymbol{x}_t)$$
(132)

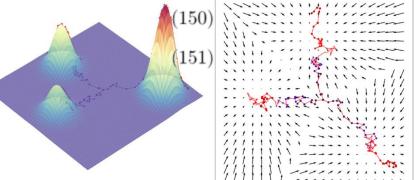
$$\therefore \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1 - \bar{\alpha}_t) \nabla \log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}}$$
 (133)

$$\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t + (1 - \bar{\alpha}_t)\nabla\log p(\boldsymbol{x}_t)}{\sqrt{\bar{\alpha}_t}} = \frac{\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\varepsilon_t}}{\sqrt{\bar{\alpha}_t}}$$

$$\therefore (1 - \bar{\alpha}_t) \nabla \log p(\mathbf{x}_t) = -\sqrt{1 - \bar{\alpha}_t} \varepsilon_t$$

$$\therefore (1 - \bar{\alpha}_t) \nabla \log p(\boldsymbol{x}_t) = -\sqrt{1 - \bar{\alpha}_t} \boldsymbol{\varepsilon}_t$$
$$\nabla \log p(\boldsymbol{x}_t) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\varepsilon}_t$$

- What is **Tweedie's Formula**?
- Conclusion:  $oldsymbol{arepsilon}_t = -\sqrt{1-ar{lpha}_t} 
  abla \log p(x_t)$



(149)

## Score Function Tv

Tweedie's Formula 补充

Tweedie's Formula 说明: 后验均值 (posterior mean) 可以通过观测值加上噪声方差乘以观测值的对数概率密度的梯度来计算。

$$egin{aligned} x_t &= \sqrt{ar{lpha}_t} x_0 + \sqrt{1 - ar{lpha}_t} arepsilon, \quad arepsilon \sim \mathcal{N}(0, 1) \ p(x_t | x_0) &= \mathcal{N}(\sqrt{ar{lpha}_t} x_0, (1 - ar{lpha}_t) \mathbf{I}) \ p(x_t | x_0) &= rac{1}{(2\pi(1 - ar{lpha}_t))^{d/2}} \mathrm{exp}(-rac{\left|\left|x_t - \sqrt{ar{lpha}_t} x_0
ight|
ight|^2}{2(1 - ar{lpha}_t)}) \end{aligned}$$

对原始图像  $x_0$  的后验和后验均值

$$egin{align} p(x_0|x_t) &= rac{p(x_t|x_0)p(x_0)}{p(x_t)} \ \mathbb{E}[x_0|x_t] &= \int x_0p(x_0|x_t)dx_0 &= rac{1}{p(x_t)}\int x_0p(x_t|x_0)p(x_0)dx_0 &= rac{1}{p(x_t)}\int x_0p(x_t|x_0)dx_0 &= rac{1}{p(x_t)}\int x_0p(x_0|x_0)dx_0 &= rac{1}{p(x_t)}\int x_0p(x_t|x_0)dx_0 &= \frac{1}{p(x_t)}\int x_0p(x_t|x_0)d$$

$$\mathbb{E}[x_0|x_t] = rac{1}{\sqrt{ar{lpha}_t}}(x_t + (1-ar{lpha}_t)
abla \log p(x_t))$$

$$-\hat{x}_0 = rac{1}{\sqrt{ar{lpha}_t}}(x_t - \sqrt{1-ar{lpha}_t}arepsilon_{ heta}(x_t,t)).$$

做一点小的推导:

$$p(x_t) = \int p(x_t|x_0) p(x_0) dx_0.$$

$$abla p(x_t) = \int 
abla p(x_t|x_0) p(x_0) dx_0$$

$$abla p(x_t|x_0) = -rac{x_t - \sqrt{ar{lpha}_t}x_0}{1-ar{lpha}_t}p(x_t|x_0)dx_0.$$

$$abla p(x_t) = \int -rac{x_t - \sqrt{arlpha_t} x_0}{1 - arlpha_t} p(x_t|x_0) p(x_0) dx_0.$$

$$abla \log p(x_t) = rac{1}{p(x_t)} 
abla p(x_t)$$

$$extstylength 
abla \log p(x_t) = rac{1}{p(x_t)} \int -rac{x_t - \sqrt{arlpha_t} x_0}{1 - arlpha_t} p(x_t|x_0) p(x_0) dx_0.$$

# Guidance Two ways to inject condition

• Way 1: Classifier-Guidance: Use an unconditional generative model  $p_{\theta}(x_{t-1}|x_t)$  (已经训练好的)

+ Classifier 
$$p_{\phi}(y|x_t)$$

Injecting Condition y in the reverse process

$$dx = \left(oldsymbol{f}_t(x) - rac{1}{2}(g_t^2 + \sigma_t^2)
abla_x \log p_t(x)
ight)dt + \sigma_t doldsymbol{w}$$

$$egin{aligned} 
abla_{x_t} \log p\left(x_t \mid y
ight) &= 
abla \log \left(rac{p\left(x_t
ight)p_\phi\left(y \mid x_t
ight)}{p(y)}
ight) \ &= 
abla \log p\left(x_t
ight) + 
abla \log p_\phi\left(y \mid x_t
ight) - 
abla \log p(y) \ &= 
abla \log p\left(x_t
ight) + 
abla \log p_\phi\left(y \mid x_t
ight) \ &= 
abla \log p\left(x_t
ight) + 
abla \log p_\phi\left(y \mid x_t
ight) \ &= 
abla \log p_\phi\left(y \mid x_t
ig$$

$$oldsymbol{arepsilon}_t = -\sqrt{1-ar{lpha}_t} 
abla \log p(x_t)$$

$$\hat{oldsymbol{arepsilon}}(x_t,t) := oldsymbol{arepsilon}_{ heta}(x_t,t) - oldsymbol{\gamma}\!\!\sqrt{1-ar{lpha}_t}
abla_{x_t}\log p_\phi(y|x_t)$$

• 注意,只改变采样过程,相当于对梯度做一个可控的偏移

# Guidance Two ways to inject condition

### **\*\*** Way 2: Classifier-Free Guidance (CFG)

直接改变训练过程  $p_{\theta}(x_{t-1}|x_t,y)$ ,  $y = label or \emptyset$ 

### Algorithm 1 Joint training a diffusion model with classifier-free guidance

**Require:**  $p_{uncond}$ : probability of unconditional training

- 1: repeat
- $(\mathbf{x}, \mathbf{c}) \sim p(\mathbf{x}, \mathbf{c})$ ▶ Sample data with conditioning from the dataset
- $\mathbf{c} \leftarrow \emptyset$  with probability  $p_{\text{uncond}} \triangleright \text{Randomly discard conditioning to train unconditionally}$
- $\lambda \sim p(\lambda)$
- Sample log SNR value  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- $\mathbf{z}_{\lambda} = \alpha_{\lambda} \mathbf{x} + \sigma_{\lambda} \boldsymbol{\epsilon}$
- Take gradient step on  $\nabla_{\theta} \| \boldsymbol{\epsilon}_{\theta}(\mathbf{z}_{\lambda}, \mathbf{c}) \boldsymbol{\epsilon} \|^2$
- > Optimization of denoising model

8: until converged

$$egin{align*} \mathbf{Sampling} & \hat{oldsymbol{arepsilon}}(x_t,t,y) := oldsymbol{arepsilon}_{ heta}(x_t,t,y) - \gamma \sqrt{1-ar{lpha}_t} 
abla_{x_t} \log p_\phi(y|x_t) \ &= oldsymbol{arepsilon}_{ heta}(x_t,t,y) - \gamma \sqrt{1-ar{lpha}_t}(
abla_{x_t} \log p_\phi(x_t|y) - 
abla_{x_t} \log p_\phi(x_t)) \ &= oldsymbol{arepsilon}_{ heta}(x_t,t,y) + \gamma (oldsymbol{arepsilon}_{ heta}(x_t,t,y) - oldsymbol{arepsilon}_{ heta}(x_t,t,y)) \ &= (1+\gamma)oldsymbol{arepsilon}_{ heta}(x_t,t,y) - \gamma oldsymbol{arepsilon}_{ heta}(x_t,t,y) \end{aligned}$$

采样时通过有条件和无条件两 种形式做一个线性外推, 用引 导系数调节控制程度

 $oldsymbol{arepsilon}_{ heta}\left(x_{t},t,y
ight) ext{ or } oldsymbol{arepsilon}_{ heta}\left(x_{t},t,\emptyset
ight)$ 

## Timeline (1)

### https://arxiv.org/pdf/1503.03585

#### Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein Stanford University

University of California, Berkeley

Niru Maheswaranatha Surya Ganguli Stanford University

A central problem in machine learning involves A central prootern in machine learning involves modeling complex data-sets using highly flexi-ble families of probability distributions in which learning, sampling, inference, and evaluation are still analytically or computationally tractable. Here, we develop an approach that simultane ously achieves both flexibility and tractability The essential idea, inspired by non-equilibriur statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. This ap-proach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of layers or time steps as well as to compute conditional and posterior probabilities under the learned model. We additionally release an open source reference imple-

#### 1. Introduction

Historically, probabilistic models suffer from a tradeoff between two conflicting objectives: tractability and flexibil-ity. Models that are tractable can be analytically evaluated and easily fit to data (e.g. a Gaussian or Laplace). However,

Proceedings of the 32<sup>nd</sup> International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copy-right 2015 by the author(s)

these models are unable to aptly describe structure in rich datasets. On the other hand, models that are flexible can be molded to fit structure in arbitrary data. For example, we can define models in terms of any (non-negative) function  $\phi(\mathbf{x})$  yielding the flexible distribution  $p(\mathbf{x}) = \frac{\phi(\mathbf{x})}{2}$ , where Z is a normalization constant. However computing this normalization constant is generally intractable. Evaluating, training, or drawing samples from such flexible models typically requires a very expensive Monte Carlo process.

JASCHA@STANFORD.EDU

NIRUM@STANFORD.EDU

SGANGULI@STANFORD.EDU

20

A variety of analytic approximations exist which amelio-rate, but do not remove, this tradeoff-for instance mean field theory and its expansions (T, 1982; Tanaka, 1998), variational Bayes (Jordan et al., 1999), contrastive divergence (Welling & Hinton, 2002; Hinton, 2002), minimum probability flow (Sohl-Dickstein et al., 2011b;a), minimum pronounty now communication (Lyu, 2011), proper scoring rules (Gneiting & Raftery, 2007; Parry et al., 2012), score matching (Hyvarinen, 2005), pseudolikelihood (Bessag, 1975), loopy belief propagation (Murphy et al., 1999), and many, many more. Non-parametric methods (Gershman & Blei, 2012)

#### 1.1 Diffusion probabilistic models

We present a novel way to define probabilistic models that

- 2. exact sampling,

https://arxiv.org/pdf/1907.05600V1

#### Generative Modeling by Estimating Gradients of the **Data Distribution**

Yang Song Stanford Universit yangsong@cs.stanford.edu

Stefano Ermon Stanford University ermon@cs.stanford.edu

#### Abstract

We introduce a new generative model where samples are produced via Langevin dynamics using gradients of the data distribution estimated with score matching. Because gradients signify he III defined when the data resides to not-dimensional regions of the control of the period data distribution for all noise becks. For sampling, we propose an annealed Langevin dynamics where we use gradients corresponding to gradually decreasing noise levels as the sampling process gets close to the data manifold. Our framework allows flexible model architectures, requires no sampling during training or for microlied model communitions. Our models not the control of the contro the use of adversarian memous, and provides a learning objective time can be used for principled model comparisons. Our models produce samples comparable to GANs on MNIST, CelebA and CIFAR-10 datasets, achieving a new state-of-the-art inception score of 8.91 on CIFAR-10. Additionally, we demonstrate that our models learn effective representations via image inpainting experiments.

#### 1 Introduction

Generative models have many applications in machine learning. To list a few, they have been used to generate high-fidelity images [22: 41, synthesize realistic speech and mains fragments [47], improve the performance of semi-supervised learning [24: 51, decid a relaxative campets and other antomalous performance of semi-supervised learning [24: 51, decid a relaxative campet and other antomalous progress is mainly driven by two approaches: likelihood-based methods [14, 25: 9, 49] and generative adversarial netwooks (GAN [13]). The foremer uses log kilcibiod (or a suitable surrogate) as the training objective, while the latter uses adversarial netwood in additional control of the design of the suitable surrogate) as the training to minimize f-divergences [34] or integral probability metrics [2, 45] between model and data distribution and the distribution of the suitable surrogate and the suitable surrogate and the surr

Although likelihood-based models and GANs have achieved great success, they have some intrinsic limitations. For example, likelihood-based models either have to use specialized architectures to build a normalized probability model (e.g., autoregressive models, flow models), or use surrogate losses (e.g., the evidence lower bound used in variational auto-encoders [25], contrastive divergence losses (e.g., the evidence fower boundard in An available allowed in Cap., the evidence for the control training (GANs avoid autorated as of the limitation of the control of the co

In this paper, we explore a new principle for generative modeling based on estimating the (Stein) score [29] of the data density, which is the gradient of the log-density function with respect to the input dimensions. This is a vector field pointing in the direction where the log data density grows the most. We use a neural network trained with some matching [27] to learn this vector field from data. We then produce samples using Langevin dynamics, which approximately works by gradually

**SGM** 2019.07 https://arxiv.org/pdf/2006.11239V1

#### **Denoising Diffusion Probabilistic Models**

Jonathan Ho UC Berkeley Ajay Jain UC Berkeley UC Berkeley

We present high quality image synthesis results using diffusion probabilistic models a class of latent variable models inspired by considerations from nonequilibrium a gass or auchi varianie musch inspired oy consisterations strom inocquiardiani in thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilism odels and denoising socre machine; with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted until progressive durfer greater decoding. On the unconditional CIPARO dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. Or 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at https://github.com/hojonathanho/diffusion.

2020

Deep generative models of all kinds have recently exhibited high quality samples in a wide variet Deep generative models of air lambs have recently extinited ingling quanty samples in a wine variety of data modalities. Generative adversarial networks (GANs), autoregressive models, flows, and variational autoencoders (VAEs) have synthesized striking image and audio samples [12, 25, 3, 55, 25, 23, 13, 30, 30, 41, 54, 24, 31, 42], and there have been remarkable advances in energy-based modeling and score matching that have produced images comparable to those of GANs [11, 52].



Figure 1: Generated samples on CelebA-HO 256 × 256 (left) and unconditional CIFAR 10 (right)

**DDPM** 2020.06

### https://arxiv.org/pdf/2011.13456v1

Preprint. Work in progress.

#### SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS

Diederik P. Kingma Google Brain abhishk@google.com ermon@cs.stanford.edu pooleb@google.com

#### ABSTRACT

Creating noise from data is easy; creating data from noise is generative modeling. We present a stochastic differential equation (SDE) that smoothly transforms a complex data distribution to a known prior distribution by slowly injecting noise, and a corresponding reverse-time SDE that transforms the prior distribution back into the data distribution by slowly removing the noise. Crucially, the reverse-time to the contraction of the contraction o into the data distribution by slowly removing the noise. Crucially, the reverse-tion SDE depends only on the time-dependent gradient field (d. A.s., score) of the control continuous generation with an uncontinuous mode, as we cannonstrain resperiments on class-conditional generation, image inpainting, and colorization. Combined with multiple architectural improvements, we achieve record-breaking performance for unconditional image generation on CITAR-1 ovi with an Inception score of 989 and FID of 2.00, a competitive likelihood of 3.10 bits/dim, and demonstrate high fidelity generation of 1024 × 1024 images for the first time from

#### 1 Introduction

Two successful classes of probabilistic generative models involve sequentially corrupting training data with slowly increasing noise, and then learning to reverse this corruption in order to form a generative model of the data. Score matching with Langevin dynamics (SMLD) (Song & Ermon, 2019) estimates the score (i.e., the gradient of the log probability density) at each noise scale, and 2019) estimates the score (i.e., the gradient of the top profosmity density) at each touse scale, and the then used Langevin dynamics to sample from a sequence of decreasing noise scales during generation. Denoising diffusion probabilistic modeling (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020) trains a sequence of probabilistic models to reverse each step of the noise corruption, using knowledge of the functional form of the reverse distributions to make training tractable. For continuous state spaces, the DDPM training objective implicitly computes scores at each noise scale. We therefore refer to these two model classes together as score-based generative models.

Score-based generative models, and related techniques (Bordes et al., 2017; Goyal et al., 2017), have proven effective at generation of images (Song & Ermon, 2019; 2020; Ho et al., 2020), audio (Chen et al., 2020; Kong et al., 2020; graphs (Niu et al., 2020; Mowever, the

**SDE** 2020.11

**Physics Foundation** 2015.11

## Timeline (2)

### https://arxiv.org/pdf/2010.02502v1

#### DENOISING DIFFUSION IMPLICIT MODELS

Jiaming Song, Chenlin Meng & Stefano Ermon Stanford University

{tsong, chenlinm, ermon}@cs.stanford.edu

Denoising diffusion probabilistic models (DDPMs) have achieved high quality image generation without adversarial training, yet the proquire similaring and Marlov chain for many gets to produce a sample. To accelerate snappling, we will be a support of the property of th

#### 1 INTRODUCTION

Deep generative models have demonstrated the ability to produce high quality samples in many domains (Germe et al., 2002; on sen Good et al., 2016a). In serme of image generation, generative adversarial revents (GANC, Goodfiller) and a (Gardy) control quality duality explaint quality gravitive models (son den Good et al., 2016b) and normalizing flows (Gezenek & Mohamed, 2015b) that al., 2016b, However, GANs require very specific thoches in optimization and architectures in order to stabilize training (Ayrowsky et al., 2015; Gardynian et al., 2017; Karrise et al., 2018; Bruck et al., 2018b) and fall as tover models of the data distribution (Januari et al., 2018b).

Construction of the constr (19) or obtained by reversing a forward diffusion process that progressively turns an image into ise (Sohl-Dickstein et al., 2015).

note (point-bettered as, M). A critical directiveske of these models is that they require many iterations to produce a high quality sample. For DDPMs, this is because that the generative process (from noise to dault approximate) and the production of the produc

To close this efficiency gap between DDPMs and GANs, we present denoising diffusion implicit models (DDIMs). DDIMs are implicit probabilistic models (Mohamed & Lakshminarayann, 2016) and are closely related to DDPMs, in the sense that they are trained with the same objective function. In Section 3, we generalize the forward diffusion process used by DDPMs, which is Markovian,

### **DDIM** 2020.10

### https://arxiv.org/pdf/2102.09672

#### Improved Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPM) are a class of generative models which have recently been shown to produce excellent sam ples. We show that with a few simple modifiations, DDPMs can also achieve competitive loglikelihoods while maintaining high sample quality Additionally, we find that learning variances of the reverse diffusion process allows sampling with an order of magnitude fewer forward passes with a negligible difference in sample quality, which is important for the practical deployment of these models. We additionally use precision and recall to compare how well DDPMs and GANs cover the target distribution. Finally, we show that the sample quality and likelihood of these models scale smoothly with model capacity and training compute, making them easily scalable. We release our code at https://github.com

#### 1. Introduction

Sohl-Dickstein et al. (2015) introduced diffusion probabilis tic models, a class of generative models which match a data distribution by learning to reverse a gradual, multi-step noising process. More recently, Ho et al. (2020) showed an equivalence between denoising diffusion probabilistic models (DDPM) and score based generative models (Song & Ermon, 2019; 2020), which learn a gradient of the logdensity of the data distribution using denoising score match-ing (Hyvärinen, 2005). It has recently been shown that this class of models can produce high-quality images (Ho et al., 2020; Song & Ermon, 2020; Jolicoeur-Martineau et al., 2020) and audio (Chen et al., 2020b; Kong et al., 2020), but it has yet to be shown that DDPMs can achieve loglikelihoods competitive with other likelihood-based models such as autoregressive models (van den Oord et al., 2016c) and VAEs (Kingma & Welling, 2013). This raises various questions, such as whether DDPMs are capable of capturing

(2020) showed extremely good results on the CIFAR-10 (Krizhevsky, 2009) and LSUN (Yu et al., 2015) datasets, it is unclear how well DDPMs scale to datasets with higher diversity such as ImageNet, Finally, while Chen et al. (2020b) found that DDPMs can efficiently generate audio using a small number of sampling steps, it has yet to be shown that the same is true for images.

In this paper, we show that DDPMs can achieve loglikelihoods competitive with other likelihood-based models tightly ontimise the variational lower-hound (VLR) we learn the reverse process variances using a simple reparame terization and a hybrid learning objective that combines the

We find surprisingly that, with our hybrid objective, our models obtain better log-likelihoods than those obtained by optimizing the log-likelihood directly, and discover that the latter objective has much more gradient noise during training. We show that a simple importance sampling tech-nique reduces this noise and allows us to achieve better log-likelihoods than with the hybrid objective.

After incorporating learned variances into our model, we surprisingly discovered that we could sample in fewer steps from our models with very little change in sample quality. While DDPM (Ho et al., 2020) requires hundreds of forward passes to produce good samples, we can achieve good samples with as few as 50 forward passes, thus speeding up sampling for use in practical applications. In parallel to our work, Song et al. (2020a) develops a different approach to fast sampling, and we compare against their approach

While likelihood is a good metric to compare against othe likelihood-based models, we also wanted to compare the the improved precision and recall metrics (Kynkäänniem et al., 2019) and discover that diffusion models achieve much higher recall for similar FID, suggesting that they do indeed cover a much larger portion of the target distribution

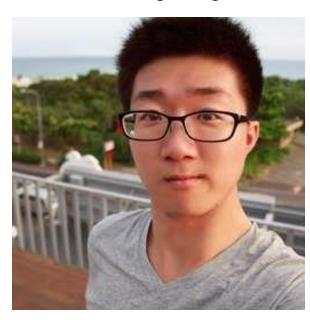
Finally, since we expect machine learning models to conall the modes of a distribution. Furthermore, while Ho et al. sume more computational resources in the future, we evalu-\*Equal contribution <sup>1</sup>OpenAI, San Francisco, USA. Correspon. ate the performance of these models as we increase mode size and training compute. Similar to (Henighan et al.,

### **Improved-diffusion** 2021.02

### Yang Song



### Jiaming Song



- 2012-2016 Tsinghua University
- 2016 PhD at Stanford University, supervised by Stefano Ermon
- OpenAI

**NVIDIA** 

## Timeline (3)

https://arxiv.org/pdf/2105.05233v1

### Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal\* OpenAI prafulla@openai.com

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 conditional models of the state of the state of the current state of the state of t

arXiv:2105.05233v1



Figure 1: Selected samples from our best ImageNet 512 × 512 model (FID 3.85

Over the past few years, generative models have gained the ability to generate human-like natural language [6], infinite high-quality synthetic images [5, 22, 44] and highly diverse human speech and usuic [57, 12]. These models can be used in a variety of ways, such a generating images from text prompts [63, 43] or learning useful feature representations [13, 7]. While these models are already

**Classifier-Guidance** 2021.05

https://arxiv.org/pdf/2207.12598

#### CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans {ionathanho,salimans}@google.com

Classifier guidance is a recently introduced method to trade off mode coverage and sample fidelity in conditional diffusion models spot training, in the same updit (Classifier guidance combines the score estimate of a diffusion model with the gradient of an image classifier and thereby requires training an image classifier spectra from the diffusion model. It also raise the question of whether guidance spectra from the diffusion model, it also raise the question of whether guidance performed by a pure generative model without such a classifier in what we call classifier free guidance, we jeintly write as conditional and sunconditional diffusion model, and we combine the resulting conditional and unconditional score obtained using classifier guidance. obtained using classifier guidance.

#### 1 INTRODUCTION

Diffusion models have recently emerged as an expressive and flexible family of generative models delivering competitive sample quality and likelihood coics on image and analos symhesis tasks (Self-Dickstein et al. 2015; Song & Ermon, 2016; Bot est al., 2020; Song et al. 2011; Kingman et al., 2021; Song et al., 2021; All: These models when delivered analos synthesis performance rivaling the quality of anexegrencies models with substantially lever inference supp (Clean et al., 2015; Kong et al., et al., 2019) and VQ-VAE-2 (Bazavi et al., 2019) in terms of FID score and classification accuracy score (16 et al., 2015; Damerand & Nicholo, 2021).

Score (1so 4.1., 2017; Insurance and Rental). A continue to boost the sample quality of a diffusion model using an extra trained classifier. Prior to classifier guidance, it was not known how to generate "bore interpretare" analyses from a diffusion model smill are to been produced by truncated Big/GAM (Brock et al., 2019) or low temperature "among the major discharge the produced by truncated Big/GAM (Brock et al., 2019) or low temperature Glow (Kingma & Dharwin), 2015; marrier attempts, such a scaling he model cost eventors of determing the amount of Genstain notice that the scaling of the scale of the control of the scale of the control of



Figure 1: Classifier-free guidance on the malamute class for a 64x64 ImageNet diffusion model. Left A short version of this paper appeared in the NeurIPS 2021 Workshop on Deep Generative Models and ownstream Applications: https://openreview.net/pdf?id=qw8&Kxf7bI https://arxiv.org/pdf/2112.10741v1

#### GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

Alex Nichol " Prafulla Dhariwal " Aditya Ramesh " Pranav Shyam Pamela Mishkin Bob McGrew Ilya Sutskever Mark Chen

#### Abstract

Diffusion models have recently been shown to generate high-quality synthetic images, especially when paired with a guidance technique to trade off diversity for fidelity. We explore diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance. We find that the latter is preferred by human evaluators for both photorealism and cap tion similarity, and often produces photorealistic samples. Samples from a 3.5 billion parameter text-conditional diffusion model using classifierfree guidance are favored by human evaluators to those from DALL-E, even when the latter uses expensive CLIP reranking. Additionally, we find that our models can be fine-tuned to perform image inpainting, enabling powerful text-driven image editing. We train a smaller model on a filtered dataset and release the code and weights at https://github.com/openai/glide-text2im.

#### 1. Introduction

Images, such as illustrations, paintings, and photographs. can often be easily described using text, but can require specialized skills and hours of labor to create. Therefore, a tool capable of generating realistic images from natural language can empower humans to create rich and diverse visual content with unprecedented ease. The ability to edit images using natural language further allows for iterative refinement and fine-grained control, both of which are critical for real world applications.

thesizing images from free-form text prompts, and can compose unrelated objects in semantically plausible ways (Xu et al., 2017; Zhu et al., 2019; Tao et al., 2020; Ramesh et al., 2021; Zhang et al., 2021). However, they are not yet able 2021; Zhang et al., 2021). However, mey are not yes some to generate photorealistic images that capture all aspects of ated for photorealism, and 69% of the time when evaluated

"Equal contribution. Correspondence to alex@openai.com, prafulla@openai.com, aramesh@openai.com

their corresponding text prompts.

On the other hand, unconditional image models can synthesize photorealistic images (Brock et al., 2018; Karras et al., 2019a;b; Razavi et al., 2019), sometimes with enough fidelity that humans can't distinguish them from real image (Zhou et al., 2019). Within this line of research, diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020b) have emerged as a promising family of generative models, achieving state-of-the-art sample quality on a number of image generation benchmarks (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho et al., 2021).

To achieve photorealism in the class-conditional setting, Dhariwal & Nichol (2021) augmented diffusion models with classifier guidance, a technique which allows diffusion models to condition on a classifier's labels. The classifier is first trained on noisy images, and during the diffusion sampling process, gradients from the classifier are used to guide the sample towards the label. Ho & Salimans (2021) achieved similar results without a separately trained classifier through the use of classifier-free guidance, a form of guidance that interpolates between predictions from a diffusion model with and without labels

Motivated by the ability of guided diffusion models to generate photorealistic samples and the ability of text-to-image models to handle free-form prompts, we apply guided dif fusion to the problem of text-conditional image synthesis. First, we train a 3.5 billion parameter diffusion model that uses a text encoder to condition on natural language descriptions. Next, we compare two techniques for guiding diffusion models towards text prompts: CLIP guidance and classifier-free guidance. Using human and automated evaluations, we find that classifier-free guidance yields higherquality images.

Recent text-conditional image models are capable of syn- We find that samples from our model generated with classifier-free guidance are both photorealistic and reflect a wide breadth of world knowledge. When evaluated by human judges, our samples are preferred to those from DALL-E (Ramesh et al., 2021) 87% of the time when evalufor caption similarity.

**CFG** 2021.12

**GLIDE** 2021.12

## Timeline (4)

https://arxiv.org/pdf/2103.00020

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford "<sup>1</sup> Jong Wook Kim <sup>1</sup> Chris Hallacy <sup>1</sup> Aditya Ramesh <sup>1</sup> Gabriel Goh <sup>1</sup> Sandhini Agarwal Girish Sastry <sup>1</sup> Amanda Askell <sup>1</sup> Pamela Mishkin <sup>1</sup> Jack Clark <sup>1</sup> Gretchen Krueger <sup>1</sup> Ilya Sutskever <sup>1</sup>

#### Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision monstrate that the simple pre-training task of predicting which caption goes with which im age is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study the performance of this approach by benchmark-ing on over 30 different existing computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet-50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

#### 1. Introduction and Motivating Work

Pre-training methods which learn directly from raw text have revolutionized NLP over the last few years (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radord et al., 2018; Devlin et al., 2018; Raffel et al., 2019).

\*Equal contribution \*OpenAI, San Francisco, CA 94110, USA. strespondence to: <{alec, jongwook}@openai.com>.

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily im proving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019: Raffel et al., 2019) has enabled taskagnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision acces sible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to pre dict the nouns and adjectives in text documents paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in captions associated with images. Sri-vastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features Joulin et al. (2016) modernized this line of work and demonstrated that CNNs trained to predict words in image cap-tions learn useful image representations. They converted the title, description, and hashtag metadata of images in the YFCC100M dataset (Thomee et al., 2016) into a bag-ofwords multi-label classification task and showed that pre training AlexNet (Krizhevsky et al., 2012) to predict these labels learned representations which preformed similarly to ImageNet-based pre-training on transfer tasks. Li et al. (2017) then extended this approach to predicting phrase ngrams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image

https://arxiv.org/pdf/2204.06125

#### **Hierarchical Text-Conditional Image Generation with CLIP Latents**

Alex Nichol Prafulla Dhariwal\* Aditya Ramesh\* Casey Chu OpenAI Mark Chen

casev@onenai.com

#### Abstract

mark@openai.com

Contrastive models like CLIP have been shown to learn robust representations of Contrastive models like CLIP have been shown to learn robust representations of image state acquire both semantics and skyl. To leverage these representations for image penetration, we propose a two-stage model: a prior that generates a CLIP image emboding given a text caption, and a decoder that generates as image representation has more timing the state of the stat

#### 1 Introduction

Recent progress in computer vision has been driven by scaling models on large datasets of captioned images collected from the internet [10, 44, 60, 39, 31, 16]. Within this framework, CLIP [39] has emerged as a successful representation learner for image. CLIP embeddings have a number of desirable properties: they are robust to image distribution shift, have impressive zero-obsc capabilities, and have been fine-tuned to achieve state-of-the-art results on a wide variety of vision and language. and have been line-tuned to achieve state-ot-the-art results on a wide variety of vision and language tasks [45]. Concurrently, diffusion models [46, 48, 25] have emerged as a promising generative modeling framework, pushing the state-of-the-art on image and video generation tasks [11, 26, 24]. To achieve best results, diffusion models leverage a guidance technique [11, 24] which improves sample fidelity (for images, photorealism) at the cost of sample diversity.

In this work, we combine these two approaches for the problem of text-conditional image generatio In this work, we combine these two approaches for the problem of net-conditional image generation. We first train additional decelerab for insert the LLPI image accessor. Our inserter is no edierministic, and can produce multiple images corresponding to a given image embedding. The presence of a present image in the problem of the presence of the contraction of the Alm Vinerion (10, 52), esconding and decoding an input image produces semantically similar output images (Figure 3). We can also interpolate between input images by inverting interpolations of their image embeddings (Figure 4). However, one mobale devaturage of using the CLIP latest space is the ability to semantically modify images by moving in the direction of any cooled set vector (Figure 5), whereas discovering these directions in GAS latest space involves.

https://arxiv.org/pdf/2205.11487

#### Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia', William Chan', Saurabh Saxena', Lala Li<sup>1</sup>, Jay Whang', Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho , David J Fleet , Mohammad Norouzi (sahariac,williamchan,mnorouzi)@google.com {srbs,lala,jwhang,jonathanho,davidfleet}@google.com Google Research, Brain Team

#### Abstract

We prevent Imagen, a text-to-image diffusion model with an unprecedented degree of photocratism and a deep level of language understanding. Imagen builds on on the strength of diffusion models in high-fidelity image generation. On the strength of diffusion models in high-fidelity image generation. On the strength of diffusion models in high-fidelity image generation. On the strength of the stren 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. See imagen.research.google for an overview of the results.

#### 1 Introduction

May 2022

23

[cs.CV]

Multimodal learning has come into prominence recently, with text-to-image synthesis [53, 12, 57] and image-text contrastive learning [40, 31, 74] at the forefront. These modes have transformed for research community and captured wheteyered applies intention with creative image generation in the properties of the prop

Imagen comprises a frozen T5-XXL [52] encoder to map input text into a sequence of embeddings and a 64×64 image diffusion model, followed by two super-resolution diffusion models for genera

**Imagen** 2022.05

### Google v.s. OpenAI



Transformer

T5

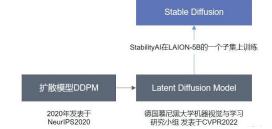
Imagen Gemini

**GPT CLIP** 

**DALLE** ChatGPT

**CLIP** 2022.03 DALLE 2 2022.04

# Timeline (5)



### https://arxiv.org/pdf/2112.10752v1

#### High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach<sup>1</sup>\* Andreas Blattmann<sup>1</sup>\* Dominik Lorenz<sup>1</sup> Patrick Esser<sup>®</sup> Björn Ommer<sup>1</sup> Ludwig Maximilian University of Munich & PWR, Heidelberg University, Germany https://github.com/CompVis/latent-diffusion

#### Abstract

nodels (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image gen-eration process without retraining. However, since these models typically operate directly in pixel space, optimizamodest specially operate airccty in paret space, opinitiza-tion of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evalu-ations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve a new state of the art for image innaint ing and highly competitive performance on various tasks, including unconditional image generation, semantic scene synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

By decomposing the image formation process into a se-quential application of denoising autoencoders, diffusion Figure 1. Boosting the upper bound on achievable quality with lent inductive biases for spatial data, we do not need the heavy spa can still greatly reduce the dimensionality of the data via suitah autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512<sup>2</sup> px. We denote the spatial downsampling factor by f. Reconstruction FIDs [26 calculated on ImageNet-val. [11]; see also Tab. 1. denoising autoencoders, have shown to achieve impressive results in image synthesis [27,80] and beyond [6,41,44,52], and define the state-of-the-art in class-conditional image

unconditional DMs can readily be applied to tasks such

is inpainting and colorization [80] or stroke-based syn-

thesis [48], in contrast to other types of generative models [17, 42, 64]. Being likelihood-based models, they do not exhibit mode-collapse and training instabilities as GANs

and, by heavily exploiting parameter sharing, they can model highly complex distributions of natural images with-out involving billions of parameters as in AR models [62].

Democratizing High-Resolution Image Synthesis DMs

mode-covering behavior makes them prone to spend ex cessive amounts of capacity (and thus compute resources

on modeling imperceptible details of the data [15,68]. Al-though the reweighted variational objective [27] aims to ad-

dress this by undersampling the initial denoising steps, DMs are still computationally demanding, since training and evaluating such a model requires repeated function evalu-

ations (and gradient computations) in the high-dimensional

space of RGB images. As an example, training the most

#### 1. Introduction

Image synthesis is one of the computer vision fields with image synthesis is one of the computer vision below with the most spectacular recent development, but also among those with the greatest computational demands. Espe-cially high-resolution synthesis of complex, natural scenes is presently dominated by scaling up likelihood-based models, potentially containing billions of parameters in autoreing results of GANs [3, 24, 36] have been revealed to be mostly confined to data with comparably limited variability as their adversarial learning procedure does not easily scale to modeling complex, multi-modal distributions. Recently, diffusion models [77], which are built from a hierarchy of

"The first two authors contributed equally to this work

https://openreview.net/pdf?id=M3 Y74vmsMcY

> LAION-5B: An open large-scale dataset for training next generation image-text models

Christoph Schuhmann §§\*\* Romain Beaumont §§\*\* Richard Vencul 3.5 §§\*\*
Cade Gordon §§\*\* Ross Wightman §§ Mehdi Cherti 3.70 §§\*
Theo Coombee's Arrush Kattle Clayton Mullis Mithed Wortsman\*
Patrick Schramowski 1.4 Srivatas Kundurthy\* Katherine Crosson 8.5 Ladvige Schmidt \*\* Robert Accurancysk 2.7 Jeins Jibres\* Jeins\* J EleutherAl9 Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)1 contact@laion.ai §§ Equal first contributions, \*\* Equal senior contributions

#### Abstract

Groundbreaking language-vision architectures like CLIP and DALL-E proved the utility of training on large amounts of noisy image-text data, without relying on expensive accurate labels used in standard vision unimodal supervised learning. The resulting models showed capabilities of strong text-guided image generation and transfer to downstream tasks, while performing remarkably at zero-shot classification with noteworthy out-of-distribution robustness. Since then, large-scale strication with noteworthy out-of-asstration robustness. Since then, targe-scale language-vision models like ALION, BASIC, GLIDE, Flamings and Imagen made further improvements. Studying the training and capabilities of such models requires datasets containing billions of image-text pairs. Until now, no datasets of this size have been made openly available for the broader research community. To address this problem and democratize research on large-scale multi-modal models, we present LAION-5B - a dataset consisting of 5.85 billion CLIP-filtered modes, we present LALUN-3B - a dataset consisting of 3.35 thinot Latt-In-inteed image-text pairs, of which 2.35 contain English language. We show successful replication and fine-tuning of foundational models like CLIP, GLIDE and Stable Diffusion using the dataset, and discuss further experiments enabled with an openly available dataset of this scale. Additionally we provide several nearest neighbor indices, an improved web-interface for dataset exploration and subset generation, and detection scores for watermark NSFW and toxic content detection

#### 1 Introduction

Learning from multimodal data such as text, images, and audio is a longstanding research challenge in machine learning [31, 51, 56, 83, 86]. Recently, contrastive loss functions combined with large neural networks have led to breakthrough in the generalization capabilities of vision and language models [58, 59, 66]. For instance, OpenA'S CLLP models [58] achieved large gains in zero-shot classification on langueNet [65], improving from the prior forp 1 accuracy of 11,58 [41] to 76.2% lan addition, CLIP achieved unprecedented performance gains on multiple challenging distribution shifts [32, 61, 67, 78, 82]. Ingerde CLIPs performance, numerous groups there further improved image-text models by increasing the amount of computation and the ratings set use [28, 28, 59, 94]. Another recent success of multimodal learning is in language generation, where DALL-E [59] and later

Project page: https://laion.ai/laion-5b-a-new-era-of-open-large-scale-multi-modal-dataset

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks

LDM v1 2021.12

LAION-5B 2022.03

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimiza tion of powerful DMs often consumes hundreds of GPL ations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. tecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve new state-of-the-art scores for im-

Abstract

https://arxiv.org/pdf/2112.10752

High-Resolution Image Synthesis with Latent Diffusion Models

https://github.com/CompVis/latent-diffusion

Dominik Lorenz<sup>1</sup> Patrick Esser® Björn Ommer<sup>1</sup>

tive biases for spatial data, we do not need the heavy spa-

tial downsampling of related generative models in latent space, but

can still greatly reduce the dimensionality of the data via suitable

can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at  $512^2$  px. We denote the spatial downsampling factor by f. Reconstruction FDs [28] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

results in image synthesis [30,85] and beyond [7,45,48,57],

and define the state-of-the-art in class-conditional image synthesis [15,31] and super-resolution [72]. Moreover, even

unconditional DMs can readily be applied to tasks such

as inpainting and colorization [85] or stroke-based syn-

els [19, 46, 69]. Being likelihood-based models, they do not

exhibit mode-collapse and training instabilities as GANs

and, by heavily exploiting parameter sharing, they can model highly complex distributions of natural images with-out involving billions of parameters as in AR models [67].

Democratizing High-Resolution Image Synthesis DMs belong to the class of likelihood-based models, whose

mode-covering behavior makes them prone to spend ex-

cessive amounts of capacity (and thus compute resources

on modeling imperceptible details of the data [16,73]. Al-

though the reweighted variational objective [30] aims to ad-

dress this by undersampling the initial denoising steps, DMs

are still computationally demanding, since training and

evaluating such a model requires repeated function evalu-ations (and gradient computations) in the high-dimensional

space of RGB images. As an example, training the most

powerful DMs often takes hundreds of GPU days (e.g. 150-1000 V100 days in [15]) and repeated evaluations on a noisy

version of the input space render also inference expensive.

thesis [53], in contrast to other types of generative mod-

Andreas Blattmann1 \*

Image synthesis is one of the computer vision fields with the most spectacular recent development, but also among those with the greatest computational demands. Espe-cially high-resolution synthesis of complex, natural scenes is presently dominated by scaling up likelihood-based mod-els, potentially containing billions of parameters in autoregressive (AR) transformers [66,67]. In contrast, the promis ing results of GANs [3, 27, 40] have been revealed to be mostly confined to data with comparably limited variability as their adversarial learning procedure does not easily scale to modeling complex, multi-modal distributions. Recently, diffusion models [82], which are built from a hierarchy of denoising autoencoders, have shown to achieve impre

age inpainting and class-conditional image synthesis and

highly competitive performance on various tasks, includ-ing text-to-image synthesis, unconditional image generation

and super-resolution, while significantly reducing computa-

tional requirements compared to pixel-based DMs.

"The first two authors contributed equally to this work

### Stable Diffusion

Stable Diffusion was made possible thanks to a collaboration with Stability AI and Runway and builds upon our previous work

https://github.com/CompVis/stable-diffusion

High-Resolution Image Synthesis with Latent Diffusion Mode



LAION, we were able to train a Latent Diffusion Model on 512x512 images from a subset of the LAION-5B database. Similar to Google's magen, this model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. With its 860M UNet and 123M text encoder, the model is relatively lightweight and runs on a GPU with at least 10GB VRAM. See this section below and the model card.

#### Requirements

A suitable conda environment named \(\text{ldm}\) can be created and activated with You can also update an existing latent diffusion environment by running pip install -e .

#### Stable Diffusion v1

Stable Diffusion v1 refers to a specific configuration of the model architecture that uses a downsampling-factor 8 autoencoder with an 860M UNet and CLIP VIT-L/14 text encoder for the diffusion model. The model was pretrained on 256x256 images and then finetuned on

Note: Stable Diffusion v1 is a general text-to-image diffusion model and therefore mirrors biases and (mis-iconceptions that are present in its training data. Details on the training procedure and data, as well as the intended use of the model can be found in the corresponding

The weights are available via the CompVis organization at Hugging Face under a license which contains specific use-based restrictions to prevent misuse and harm as informed by the model card, but otherwise remains permissive. While commercial use is permitted under the terms of the license, we do not recommend using the provided weights for services or products without additional safety mechanisms and considerations, since there are known limitations and biases of the weights, and research on safe and ethical deployment of general text-to-image models is an ongoing effort. The weights are research artifacts and should be treated as such

The CreativeML OpenRAIL M license is an Open RAIL M license, adapted from the work that BigScience and the RAIL Initiative are jointly carrying in the area of responsible Al licensing. See also the article about the BLOOM Open RAIL license on which our license is based.

LDM v2 2022.04

### Stable Diffusion V1 2022,08

## Latent Diffusion Model

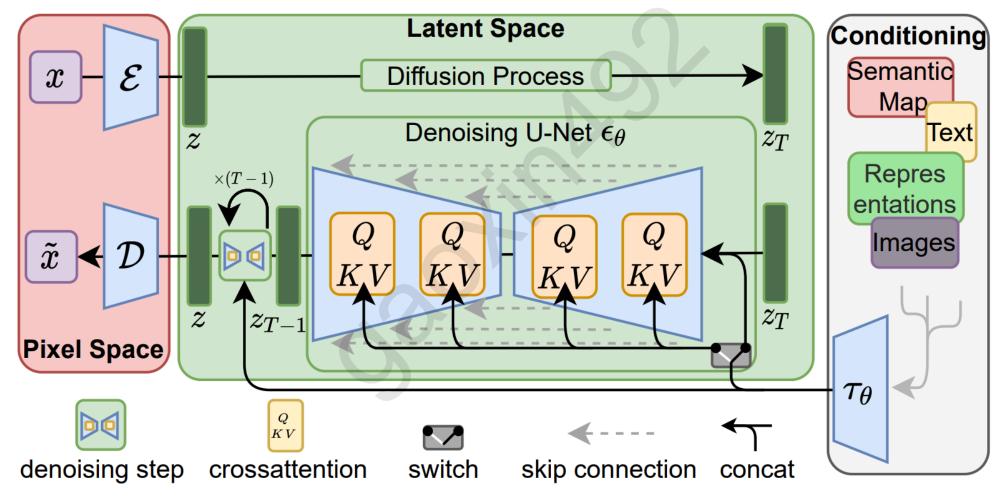






Patrick Esser

Robin Rombach



# VQ-VAE

Vector Quantized Variational AutoEncoder

$$z_q(x) = e_k$$
, where  $k = \operatorname{argmin}_j ||z_e(x) - e_j||_2$ 

• sg (stop gradient)  $L_{\mathrm{recon}} = \left|\left|x - \mathrm{decoder}(z_e(x) + sg(z_q(x) - z_e(x)))\right|\right|_2^2$ 

$$\left|L_e = \left|\left|z_e(x) - z_q(x)
ight|
ight|^2_2 \quad 
ightarrow \quad \left|L_e = \left|\left|sg(z_e(x)) - z_q(x)
ight|
ight|^2_2 + eta |\left|z_e(x) - sg(z_q(x))
ight|
ight|^2_2$$

$$\implies L = L_{\text{recon}} + \alpha L_e$$

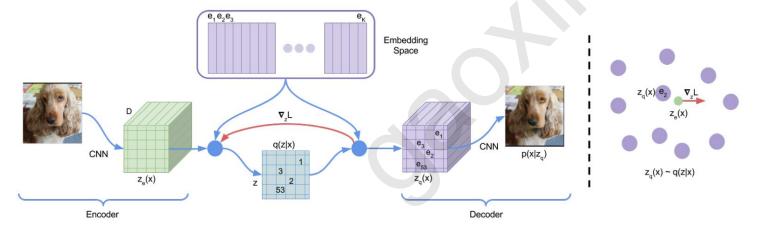


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder z(x) is mapped to the nearest point  $e_2$ . The gradient  $\nabla_z L$  (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

 $L = x - decoder(z_e + (z_q - z_e).detach())$ 

### https://arxiv.org/pdf/1711.00937

#### **Neural Discrete Representation Learning**

Aaron van den Oord DeepMind avdnoord@google.com Oriol Vinyals
DeepMind
vinyals@google.com

Koray Kavukcuoglu DeepMind korayk@google.com

#### Abstract

Learning useful representations without supervision remains a key challenge in machine learning. In this paper, we propose a simple yet powerful generative model that learns such discrete representations. Our model, the Vector Quantised Variational AutoEncoder (VQ-VAE), differs from VAEs in two key ways: the encoder network outputs discrete, rather than continuous, codes; and the prior is learnt rather than static. In order to learn a discrete latent representation, we incorporate ideas from vector quantisation (VQ). Using the VQ method allows the model to circumvent issues of "posterior collapse"—where the latents are ignored when they are paired with a powerful autoregressive decoder—typically observed in the VAE framework. Pairing these representations with an autoregressive prior the model can generate high quality images, videos, and speech as well as doing high quality speaker conversion and unsupervised learning of phonemes, providing further evidence of the utility of the learnt representations.

#### 1 Introduction

arXiv:1711.00937v2 [cs.LG] 30 May 2018

Recent advances in generative modelling of images [38, 12, 13, 22, 10], audio [37, 26] and vidcos [20, 11] have yielded impressive samples and applications [24, 18]. At the same time, challenging tasks such as few-shot learning [34], domain adaptation [17], or reinforcement learning [35] heavily rely on learnt representations from raw data, but the usefulness of generic representations trained in an unsupervised fashion is still far from being the dominant approach.

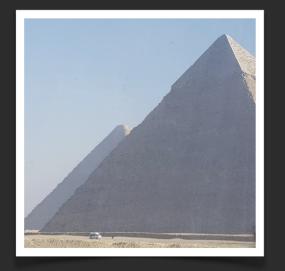
Maximum likelihood and reconstruction error are two common objectives used to train unsupervised models in the pixel domain, however their usefulness depends on the particular application the features are used in. Our goal is to achieve a model that conserves the important features of the data in its latent space while optimising for maximum likelihood. As the work in [7] suggests, the best generative models (as measured by log-likelihood) will be those without latents but a powerful decoder (such as PixelCNN). However, in this paper, we argue for learning discrete and useful latent variables, which we demonstrate on a variety of domains.

Learning representations with continuous features have been the focus of many previous work [16, 39, 6, 9] however we concentrate on discrete representations [27, 33, 8, 28] which are potentially a more natural fit for many of the modalities we are interested in. Language is inherently discrete, similarly speech is typically represented as a sequence of symbols. Images can often be described concisely by language [40]. Furthermore, discrete representations are a natural fit for complex reasoning, planning and predictive learning (e.g., if it rians, it will use an unbrella). While using discrete latent variables in deep learning has proven challenging, powerful autoregressive models have been develooed for modelling distributions over discrete variables [37].

In our work, we introduce a new family of generative models succesfully combining the variational autoencoder (VAE) framework with discrete latent representations through a novel parameterisation of the posterior distribution of (discrete) latents given an observation. Our model, which relies on evector quantization (VQ), is simple to train, does not suffer from large variance, and avoids the

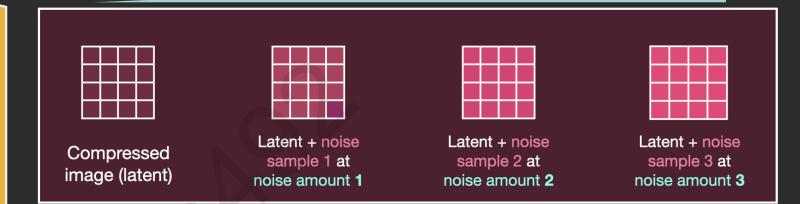
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Original image



Generate training examples with different amounts of noise added to their compressed/latent version

Image Encoder



Generated image

Image Decoder Image Generation by Reverse Diffusion (Denoising)

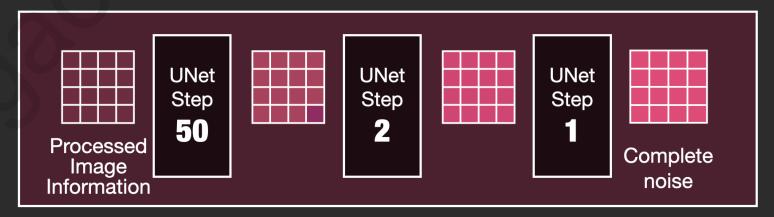


Image Information Creator

## Condition

### 1. Cross Attention in UNet

https://arxiv.org/pdf/2112.10752v1

To pre-process y from various modalities (such as language prompts) we introduce a domain specific encoder  $\tau_{\theta}$  that projects y to an intermediate representation  $\tau_{\theta}(y) \in \mathbb{R}^{M \times d_{\tau}}$ , which is then mapped to the intermediate layers of the UNet via a cross-attention layer implementing Attention $(Q,K,V)=\operatorname{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V$ , with

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), \ K = W_K^{(i)} \cdot \tau_{\theta}(y), \ V = W_V^{(i)} \cdot \tau_{\theta}(y).$$

Here,  $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$  denotes a (flattened) intermediate representation of the UNet implementing  $\epsilon_\theta$  and  $W_V^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$ ,  $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$  &  $W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$  are learnable projection matrices [32,91]. See Fig. 3 for a visual depiction.

Based on image-conditioning pairs, we then learn the conditional LDM via

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \Big[ \|\epsilon - \epsilon_{\theta}(z_t, t, \tau_{\theta}(y))\|_2^2 \Big], \quad (3)$$

where both  $\tau_{\theta}$  and  $\epsilon_{\theta}$  are jointly optimized via Eq. 3. This conditioning mechanism is flexible as  $\tau_{\theta}$  can be parameterized with domain-specific experts, *e.g.* (unmasked) transformers [91] when y are text prompts (see Sec. 4.3.1)

### 2. Different conditioning method

https://arxiv.org/pdf/2212.09748

### **Scalable Diffusion Models with Transformers**

William Peebles\*
UC Berkeley

Saining Xie
New York University

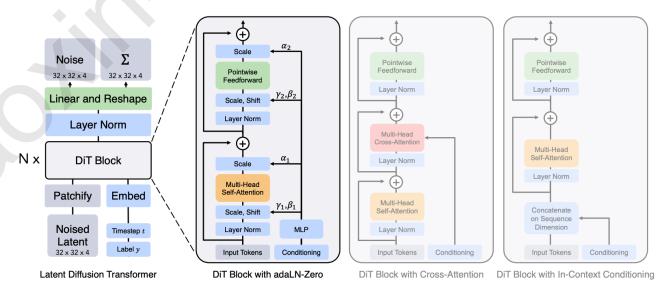
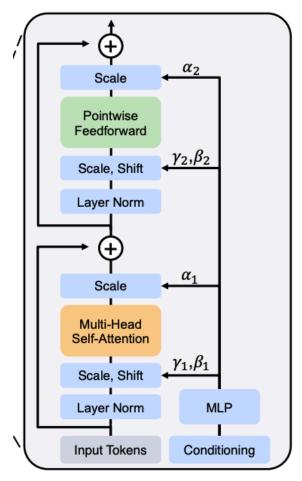


Figure 3. The Diffusion Transformer (DiT) architecture. Left: We train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks. Right: Details of our DiT blocks. We experiment with variants of standard transformer blocks that incorporate conditioning via adaptive layer norm, cross-attention and extra input tokens. Adaptive layer norm works best.

## Condition



DiT Block with adaLN-Zero

Adaptive Layer Normalization

$$y = rac{x - \mathbb{E}[x]}{\sqrt{ ext{Var}[x] + \epsilon}} \cdot oldsymbol{\gamma}(t,c) + oldsymbol{eta}(t,c)$$
AdaLN-Zero initialize  $oldsymbol{lpha} = 0$ 

$$\alpha(y) \odot f(x) + x$$

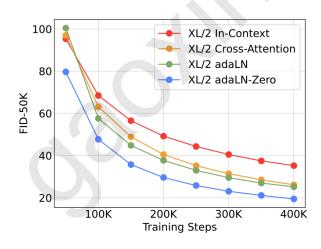
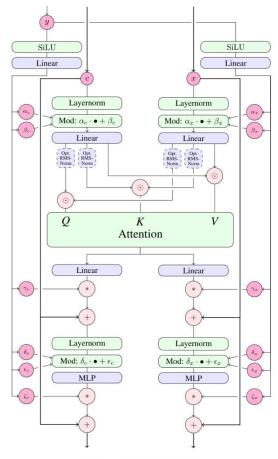


Figure 5. Comparing different conditioning strategies. adaLN-Zero outperforms cross-attention and in-context conditioning at all stages of training.

### 3. MM-DiT in Stable Diffusion v3



(b) One MM-DiT block

https://arxiv.org/abs/2403.03206

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

<sup>2</sup>NVIDIA

Text-to-image models offer unprecedented freedom to guide creation through natural lant, it is unclear how such freedom can be exercised to generate images of specific scepts, modify their appearance, or compose them in new roles and novel scenes unique concepts, modify their appearance, or compose them in new roles and movel scenes. In other words, see also been can we ten imageney guided models to term our call not a few of the control of the Our code, data and new words will be available at: https://textual-inversion















Figure 1: (left) We find new pseudo-words in the embedding space of a pre-trained text-to-image model which describe expedit concepts. (right) These pseudo-words can be compared into new sentences, placing

\* Work was done during an internship at NVIDIA

Preprint. Under review

**Textual Inversion** 2022.08

https://arxiv.org/pdf/2208.122 42v1

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

Nataniel Ruiz<sup>1,2</sup>, Yuanzhen Li<sup>1</sup>, Varun Jampani<sup>1</sup>, Yael Pritch<sup>1</sup>, Michael Rubinstein<sup>1</sup>, and Kfir Aberman<sup>1</sup>



Figure 1: With just a few images (typically 3-5) of a subject (left), DreamBooth—our AI-powered photobooth—can generate a myraid of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel articulations and vaniation in lighting conditions, all while maintaining high fidelity to the key visual features of the subject. Image credit (input images): Unsplanh.

#### Abstract

Large text-to-image models achieved a remarkable loop in the evolution of AI, enabling high-quality and diverse syndhesis of images from a given text groung. However, these models has the shill byto manitum being quarter of subjects in a given reference set and models have a subject to a price of subjects in a given reference set and proach for "perconalization" of sext-to-image diffusion models (specializing them to users' needs). Given as inspiral as few images of a subject, when them a permiande text-to-order of the subject is a few interest of the subject in the court of the subject is a few interest of the subject in the court of them and the subject is forward to the court of the subject is considered in the courty domain of the model, the image identifier can he be used to synthesize the subject in diverse scene, so the subject is deverse scene, so, so, view, and lighting conditions that done opposit in the reference images. We apply tool, text-guided very vallets, and general conditions, text-guided very vallets, and seen conditions, and artistic rendering all which contact greater the conditions, text-guided very vallets, against greater gooding the sensitive rendering all which contact greater the subject to the contact greater than the cont tion, text-guided view synthesis, appearance modification, and artistic rendering (all whi preserving the subject's key features). Project page: https://dreambooth.github.io

2022,08

https://arxiv.org/pdf/2301.07 093v1

### GLIGEN: Open-Set Grounded Text-to-Image Generation

Yuheng Li<sup>1§</sup>, Haotian Liu<sup>1§</sup>, Qingyang Wu<sup>2</sup>, Fangzhou Mu<sup>1</sup>, Jianwei Yang<sup>3</sup>, Jianfeng Gao<sup>3</sup>, Chunyuan Li<sup>3¶</sup>, Yong Jae Lee<sup>1¶</sup>

<sup>1</sup>University of Wisconsin-Madison <sup>2</sup>Columbia University <sup>3</sup>Microsoft https://gligen.github.io



Figure 1. GLIGEN enables versatile grounding capabilities for a frozen text-to-image generation model, by feeding different grounding conditions. GLIGEN supports (a) text entity + box, (b) image entity + box, (c) image style and text + box, (d) text entity + keypoints. The generated examples for each scenario are shown in ton-left, ton-right, bottom-left, and bottom-right, respectively.

#### Abstract

Large-scale text-to-image diffusion models have made amazing advances. However, the status quo is to use text input alone, which can impede controllability. In this work, we propose GLIGEN, Grounded-Language-to-Image Generation, a novel approach that builds upon and extends the functionality of existing pre-trained text-to-image diffusion models by enabling them to also be conditioned on grounding inputs. To preserve the vast concept knowledge of the pre-trained model, we freeze all of its weights and inject the grounding information into new trainable layers via a gated mechanism. Our model achieves open-world grounded text2img generation with caption and bounding box condition inputs, and the grounding ability generalizes well to novel spatial configurations and concepts. GLIGEN's zero-shot performance on COCO and LVIS outperforms existing supervised layout-to-image baselines by a large margin. § Part of the work performed at Microsoft: ¶ Co-senior authors

were the state-of-the-art, with their latent space and con-ditional inputs being well-studied for controllable manipulation [46, 58] and generation [27, 29, 45, 80]. Text conditional autoregressive [50,72] and diffusion [49,54] models have demonstrated astonishing image quality and concept coverage, due to their more stable learning objectives and large-scale training on web image-text paired data. These models have gained attention even among the general public due to their practical use cases (e.g., art design and creation).

1. Introduction

Despite exciting progress, existing large-scale text-toimage generation models cannot be conditioned on other input modalities apart from text, and thus lack the ability to precisely localize concepts or use reference images to ural language alone, restricts the way that information can

Image generation research has witnessed huge advance

in recent years. Over the past couple of years, GANs [14]

**GLIGEN** 2023.01

### https://arxiv.org/pdf/2302.05543

**Control** 

### Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala Stanford University









Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), etc., to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

We present ControlNet, a neural network architecture to add spatial conditioning controls to large, pretrained text to-image diffusion models. ControlNet locks the production ready large diffusion models, and reuses their deep and robust encoding layers pretrained with billions of images as a strong backbone to learn a diverse set of conditional control.

The neural architecture is connected with "zero convolutions (zero-initialized convolution layers) that progressively grow (zero-initialized convolution layers) that progressively grow the parameters from zero and ensure that no harmful noise could affect the finetuning. We test various conditioning con-rolls, e.g., edge, depth, sepensation, human pose, etc., with Stable Diffusion, using single or multiple conditions, with or without prompts. We show that the training of Control-Nets is robust with small (~50k) and large (>1m) datasets. Extensive results show that ControlNet may facilitate wider applications to control image diffusion models.

Many of us have experienced flashes of visual inspiration that we wish to capture in a unique image. With the advent of text-to-image diffusion models [54, 62, 72], we can now create visually stunning images by typing in a text prompt. Yet, text-to-image models are limited in the control they provide over the spatial composition of the image; precisely expressing complex layouts, poses, shapes and forms can be difficult via text prompts alone. Generating an image that accurately matches our mental imagery often requires no merous trial-and-error cycles of editing a prompt, inspecting

the resulting images and then re-editing the prompt. Can we enable finer grained spatial control by letting users provide additional images that directly specify their desired image composition? In computer vision and machine learning, these additional images (e.g., edge maps, human pose skeletons, segmentation maps, depth, normals, etc.) are often treated as conditioning on the image generation process. Image-to-image translation models [34, 38] learn https://arxiv.org/pdf/2302.084 53v1

#### T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for **Text-to-Image Diffusion Models**

Chong Mou\*1,2 Xintao Wang†2 Liangbin Xie\*2,3,4 Jian Zhang†1 Zhongang Qi<sup>2</sup> Ying Shan<sup>2</sup> Xiaohu Qie<sup>2</sup>

<sup>1</sup>Peking University Shenzhen Graduate School <sup>2</sup>ARC Lab, Tencent PCG <sup>3</sup>University of Macau <sup>4</sup>Shenzhen Institute of Advanced Technology "A doll in the shope of letter 'A'











**T2I-Adapter** 2023.02

DreamBooth

**ControlNet** 2023.02

### IC-Light

《Scaling In-the-Wild Training for Diffusion-Based Illumination Harmonization and Editing by Imposing Consistent Light Transport》

- Author: Lymin Zhang 苏大本科 => Stanford 博
- Task: Illumination harmonization and editing
- **Difficulty**: Preserving the underlying image details and maintaining intrinsic properties unchanged.
- Goal: Precise illumination manipulation

#### Adding conditional control to text-to-image diffusion models

L Zhang, A Rao, M Agrawala

Proceedings of the IEEE/CVF International Conference on ..., 2023 openaccess.thecvf.com

#### Abstract

We present ControlNet, a neural network architecture to add spatial conditioning controls to large, pretrained text-to-image diffusion models. ControlNet locks the production-ready large diffusion models, and reuses their deep and robust encoding layers pretrained with billions of images as a strong backbone to learn a diverse set of conditional controls. The neural architecture is connected with" zero convolutions (zero-initialized convolution

layers) that progressively grow the parameters from zero and ensure that no harmful noise

展开~

☆ 保存 ⑰ 引用 被引用次数: 3036 相关文章 所有 6 个版本 >>>

- Method: Impose Consistent Light (IC-Light) transport during training (rooted in physical principle)
- **Results**: Stable and scalable illumination learning, scale up the training of diffusion-based illumination editing models to large data quantities, reduces uncertainties and mitigates artifacts...

### Illumination harmonization and editing

• Typical Use Case:
Users give an object
image and illumination
description, and our
method generates
corresponding object
appearances and
backgrounds.

#### • Challenge:

- ①加上的东西 Line49
- ②本身有的东西 Line86



input

input



"... sunlight through the blinds, near window blinds"



"... magic golden lit, forest"



input



"... sunlight from the left side, beach"



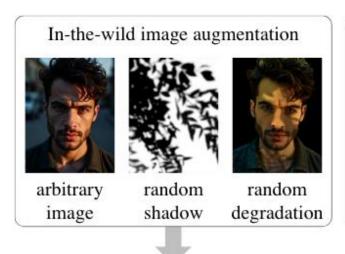
input

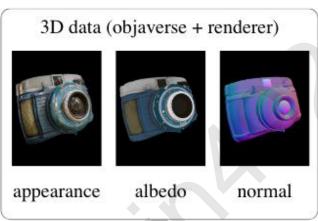


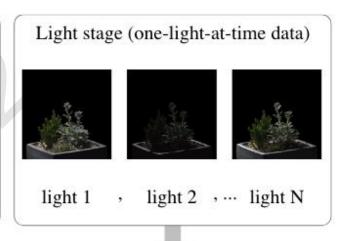
"... neo punk, city night"

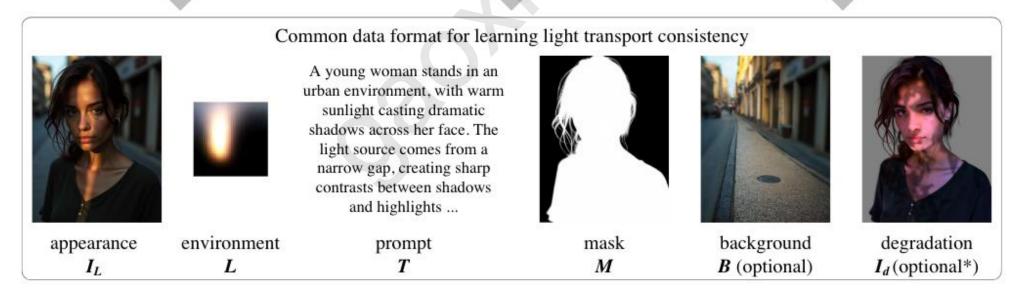
### Dataset formation

#### 用了很多别人训好的功能特异的模型来构造数据集









(a) The vanilla objective will often lead to random model behaviors, e.g., color mismatch, incorrect details, etc.

$$\mathcal{L}_{\text{vanilla}} = \|\epsilon - \delta(\varepsilon(I_L)_t, t, L, \varepsilon(I_d))\|_2^2$$

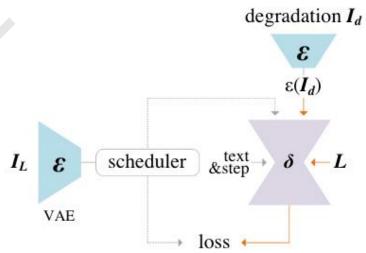
(a) In computational photography, light transport theory demonstrates that, considering arbitrary appearance  $I_L^*$  and the correlated environment illumination L, a matrix T always exists so that

$$I_L^* = TL$$

Because of this linearity, light transport explains appearance merging that

$$I_{L_1+L_2}^* = T(L_1 + L_2) = I_{L_1}^* + I_{L_2}^*$$

where  $L_1$ ,  $L_2$  are two arbitrary environment illumination maps.



(a) Vanilla image-conditioned diffusion

This intuitively shows that the mixture of an object's appearances under separate illuminations (e.g.,  $L_1$ ,  $L_2$ ) is equivalent to the appearance under merged illumination (e.g.,  $I_{L_1+L_2}^*$ ).

$$m{I}_{m{L}_1 + m{L}_2}^* = m{T}(m{L}_1 + m{L}_2) = m{I}_{m{L}_1}^* + m{I}_{m{L}_2}^*$$

This intuitively shows that the mixture of an object's appearances under separate illuminations (e.g.,  $L_1$ ,  $L_2$ ) is equivalent to the appearance under merged illumination (e.g.,  $I_{L_1+L_2}^*$ ).

怎么把这个一致性约束加到 Diffusion 的损失函数里面去?



Scene with illumination A (real photo)



Scene with illumination B (real photo)



Scene with illumination C (real photo)



Blending of real photo A and B (computed image)



Altered Blending with color tone (computed image)

Figure 1: Examples for "the linear blending of an object's appearances under different illumination conditions is consistent with its appearance under mixed illumination". Images from OToole (2016).

$$m{I}_{m{L}_1+m{L}_2}^* = m{T}(m{L}_1+m{L}_2) = m{I}_{m{L}_1}^* + m{I}_{m{L}_2}^*$$
 \*表示 images in raw high-dynamic range

#### 1. Image Space: Image => Predicted Noise 图像的线性关系可以转变为噪声的线性关系

"Clean image + Noise = Noisy Image"  $\implies$  "Estimated Clean image = Noisy Image - Predicted Noise"

A simple k-diffusion epsilon target at sigma-space step  $\sigma_t$ , estimated noise  $\epsilon_L$  (conditioned on L), and noisy image  $I_{\sigma_t}$ , the estimated clean appearance  $\hat{I_L} = (I_{\sigma_t} - \epsilon_L)/\sigma_t$ 

$$\epsilon_{\boldsymbol{L}_1+\boldsymbol{L}_2} = \epsilon_{\boldsymbol{L}_1} + \epsilon_{\boldsymbol{L}_2} \implies \|\epsilon_{\boldsymbol{L}_1+\boldsymbol{L}_2} - (\epsilon_{\boldsymbol{L}_1} + \epsilon_{\boldsymbol{L}_2})\|_2^2$$

2. Latent Space: Linear summation relation  $\Rightarrow$  MLP mapping  $\phi$ 

$$\mathcal{L}_{\text{consistency}} = \|\boldsymbol{M} \odot (\epsilon_{\boldsymbol{L}_1 + \boldsymbol{L}_2} - \phi(\epsilon_{\boldsymbol{L}_1}, \epsilon_{\boldsymbol{L}_2}))\|_2^2$$

Intuition: Assume mapping f: latent space  $\rightarrow$  image space  $f(\epsilon_{L_1+L_2}) = f(\epsilon_{L_1}) + f(\epsilon_{L_2}) \Rightarrow \epsilon_{L_1+L_2} = f^{-1}(f(\epsilon_{L_1}) + f(\epsilon_{L_2})) \Rightarrow \epsilon_{L_1+L_2} = \phi(\epsilon_{L_1}, \epsilon_{L_2})$ 

3. Implementation of environment illumination maps

$$\mathcal{L}_{\text{consistency}} = \| \boldsymbol{M} \odot (\epsilon - \phi(\boldsymbol{\delta}(\boldsymbol{\varepsilon}(\boldsymbol{I}_{\boldsymbol{L}_1})_t, t, \boldsymbol{L}_1, \boldsymbol{\varepsilon}(\boldsymbol{I}_d))), \boldsymbol{\delta}(\boldsymbol{\varepsilon}(\boldsymbol{I}_{\boldsymbol{L}_2})_t, t, \boldsymbol{L}_2, \boldsymbol{\varepsilon}(\boldsymbol{I}_d))) \|_2^2$$

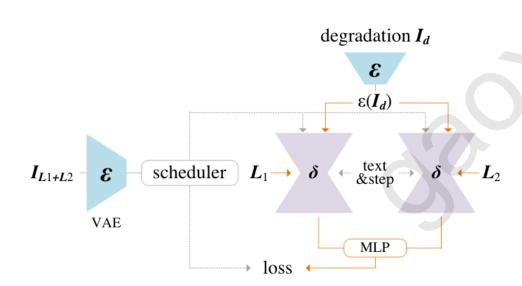
**Joint learning objective** The final learning objective can be written as

$$\mathcal{L} = \lambda_{\text{vanilla}} \mathcal{L}_{\text{vanilla}} + \lambda_{\text{consistency}} \mathcal{L}_{\text{consistency}},$$

where  $\mathcal{L}$  is the merged objective, and we use  $\lambda_{\text{vanilla}} = 1.0, \lambda_{\text{consistency}} = 0.1$  as default weights.

# 3. Implementation of environment illumination maps

满足  $L = L_1 + L_2$ 



(b) Learning light transport consistency

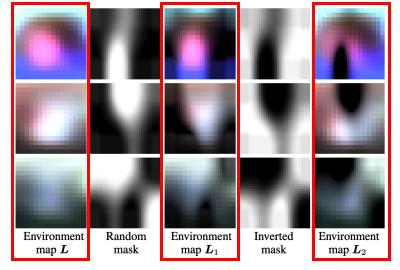


Figure 4: Examples of decomposed environment maps. We present examples to use random masks to decompose environment map  $\boldsymbol{L}$  into  $\boldsymbol{L}_1$  and  $\boldsymbol{L}_2$ . Note that  $\boldsymbol{L} = \boldsymbol{L}_1 + \boldsymbol{L}_2$ . A typical full environment map is usually of ratio 2:1, with size  $64 \times 32$  when convoluted. We use the front half (facing the image) of the convoluted environment map, which is  $32 \times 32$ . Using the front half makes normal-based environment extraction easier (since the image-space normals often do not have any pixels facing to the back half). Besides, the back halfs of environment maps from DiffusionLight Phongthawee et al. (2023) are usually not strictly correlated to image contents and can be excluded.

### Experiments

#### • Metric:

PSNR: 基于像素差异, 简单

SSIM: 通过结构信息评估图像相似度

LPIPS: 基于深度学习的感知评价

• Inference: Condition on (Image ⊙ Foreground Mask), Illumination maps + Text Prompt

Table 1: Quantitative tests of ablative architectures and alternative methods.

Method	$PSNR \uparrow$	SSIM $\uparrow$	$LPIPS \downarrow$
SwitchLight	18.45	0.7024	0.3245
DiLightNet	21.78	0.8013	0.1721
w/o LTC	20.32	0.7542	0.1927
w/o aug. data	23.95	0.8723	0.1115
w/o 3d data	22.10	0.8041	0.1298
w/o light stage	23.70	0.8501	0.1077
Ours	23.72	0.8513	0.1025

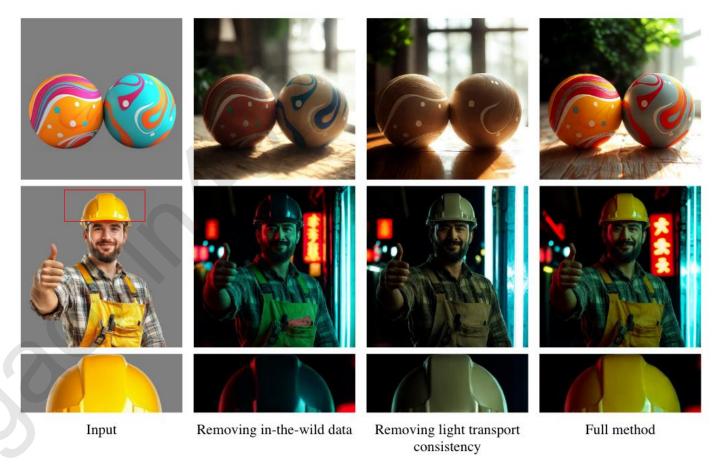
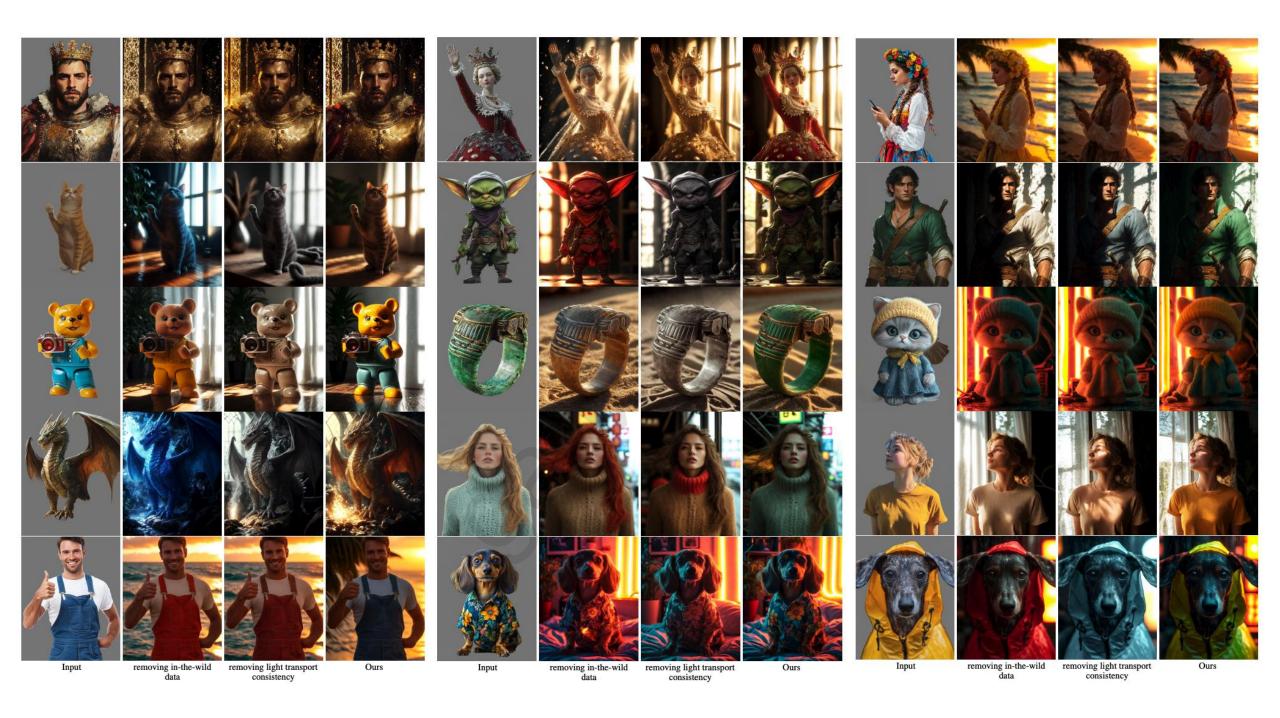


Figure 4: **Ablative Study.** We present results by removing the light transport consistency or in the wild data. More results are in the supplementary material. Results in this figure are from Stable Diffusion 1.5 version of our model. Prompts are "toy in room, studio lighting", and "a handsome man, neon city".



## Additional Application

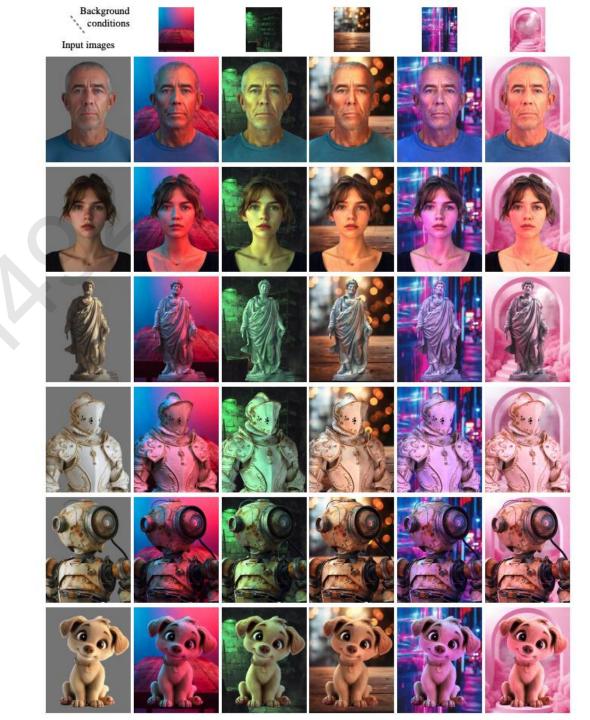
- Background-conditioned Model
- 1 Training:

"Besides, to train background-conditioned model, we concatenate B to  $I_d$  (and fill the extra channel with all zeros if some part of the dataset do not have backgrounds)."

**②** Inference:

(Image © Foreground Mask), Background conditions

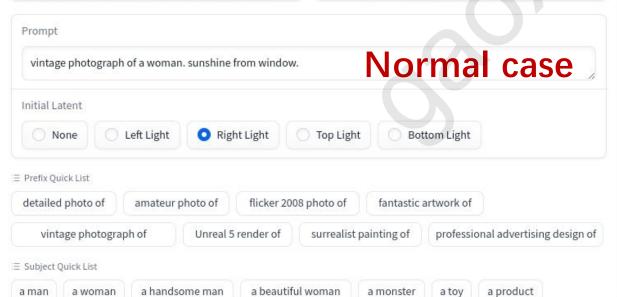
- Alternative base diffusion models SD 1.5, SDXL, Flux
- Normal Estimation (Omitted)

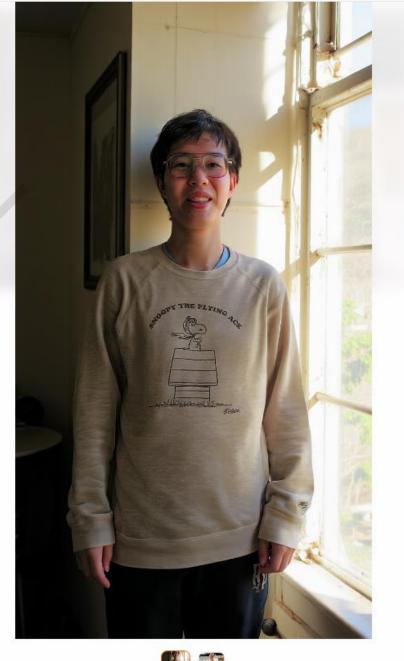


Flux-based IC-Light Model with 16ch VAE and native high resolution. See also https://github.com/lllyasviel/IC-Light/discussions/98







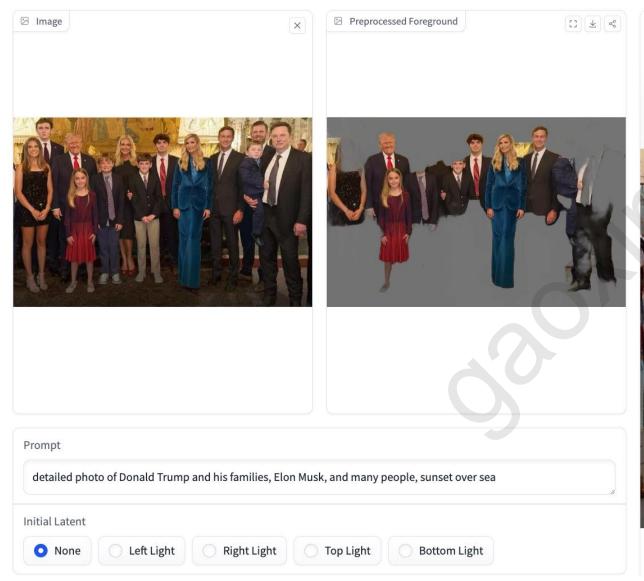


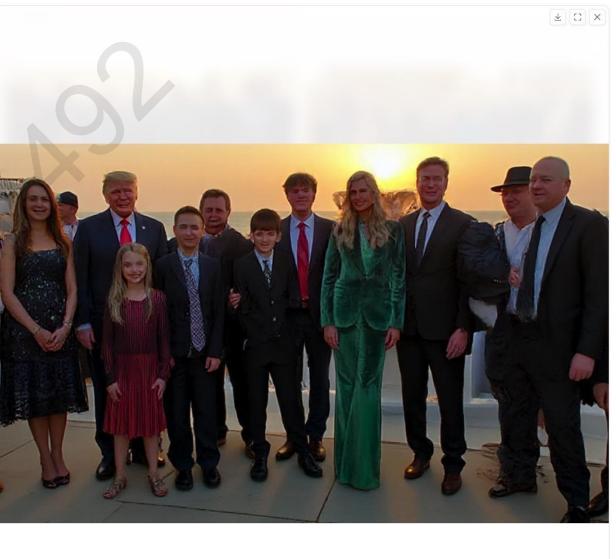


#### IC-Light V2

Flux-based IC-Light Model with 16ch VAE and native high resolution. See also https://github.com/lllyasviel/IC-Light/discussions/98

### 前景很多的 case

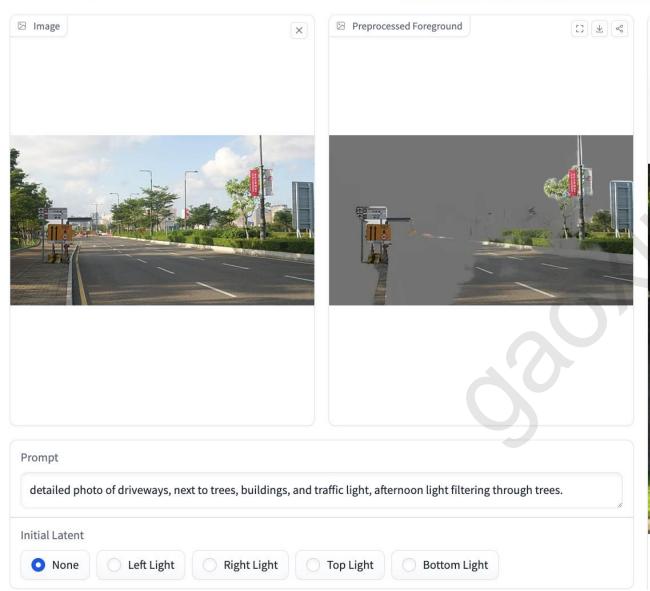


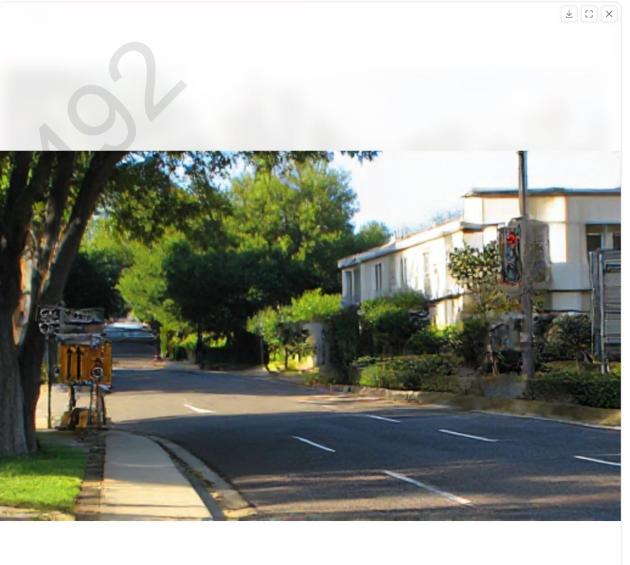


#### IC-Light V2

Flux-based IC-Light Model with 16ch VAE and native high resolution. See also https://github.com/lllyasviel/IC-Light/discussions/98







### Preview of the later lecture

- 最优扩散方差估计
- SDE and ODE
- Score-based Generative Model
- Pseudo Numerical Methods for Diffusion Models on Manifolds: PNMD/PLMS, 对DDPM的改进
- 加速采样
- Flow Matching
- Rectified Flow
- 大图生成 upscaling
- 蒸馏 for one-step generation
- Consistency Model

### References

See in main text

• Others:

[1] Luo C. Understanding diffusion models: A unified perspective. arXiv 2022[J]. arXiv preprint arXiv:2208.11970.

[2] Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2023, 56(4): 1-39.

• Other resources you may refer to:

https://github.com/Fafa-DL/Lhy\_Machine\_Learning

https://huggingface.co/docs/diffusers/index

https://jalammar.github.io/illustrated-stable-diffusion/