扩散模型：方法与应用

邓志杰

上海交通大学 清源研究院

* 部分素材取自朱军教授、李崇轩教授关于扩散模型的报告

背景：生成式AI正迅速发展

Midjourney V5创建高分辨率的逼真图像

Prompt: kids party favors

Prompt: sand and water in a desert, in the style of metropolis meets nature, les nabis, majestic ports, national geographic photo, tangled nests, dark emerald and beige, aerial view

Prompt: film still of 16 year old blond girl. looks like alice in wonderland. wearing modern clothes designed by Alexander McQueen, balancing joyously but happy. Extremely detailed and realistic. colorful garden. sharp focus. in london. Directed by Tony McNamara

https://legacy.midjourney.com/showcase/recent/

背景：生成式AI正迅速发展

DALL·E 3直接理解复杂的自然语言指令

I am doing a report on cirrus clouds for my high school science class. I need photorealistic images that show off how wispy they are. I am going to compare them to photos I took of puffy cumulonimbus clouds at my house yesterday.

Here are the photorealistic images of cirrus clouds showcasing their wispy and delicate nature. You can compare these with the photos of cumulonimbus clouds you took to highlight the differences in their structures and appearances. If you need any further information or assistance, please let me know!

My business partner and I are creating a coffee shop inspired by our mutual love for rabbits. We have a logo we like a lot that is minimalist with a rabbit, but anxious to get your take on it.

Here are some minimalist rabbit-inspired logo designs for your coffee shop. Each captures the essence of a rabbit while integrating coffee-related elements. I hope these designs align with the vision you and your business partner have for your coffee shop. If you have any feedback or further preferences, please let me know!

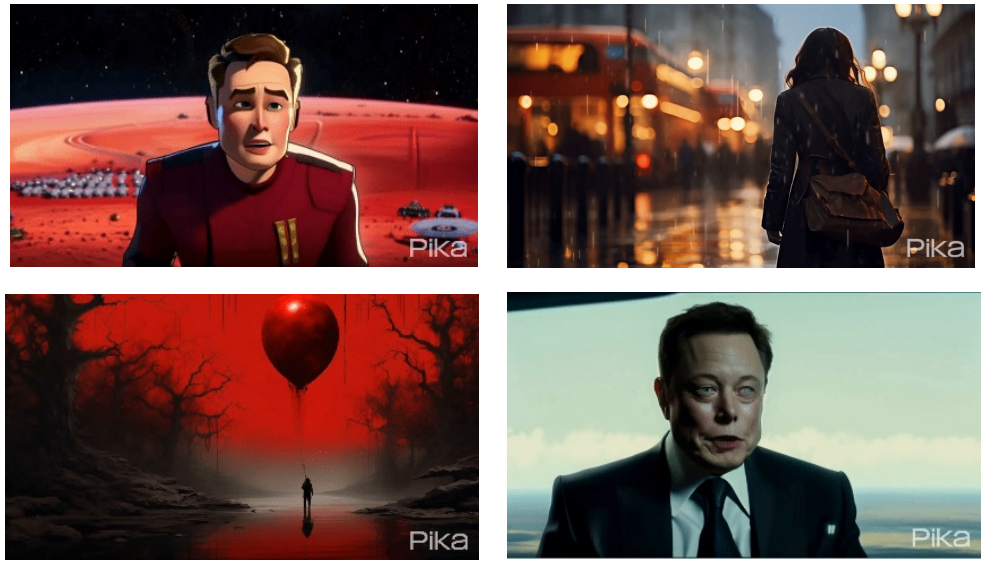
I am working with a web designer for a new website to sell bicycle helmets. Could you recommend a few website UI's that are both modern and simple that I can share with her as examples?

Here are some website UI designs tailored for selling bicycle helmets with a modern and simple aesthetic:

https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise

背景：生成式AI正迅速发展

视频生成



Pika 1.0 by Pika Labs

Gen-2 by Runway

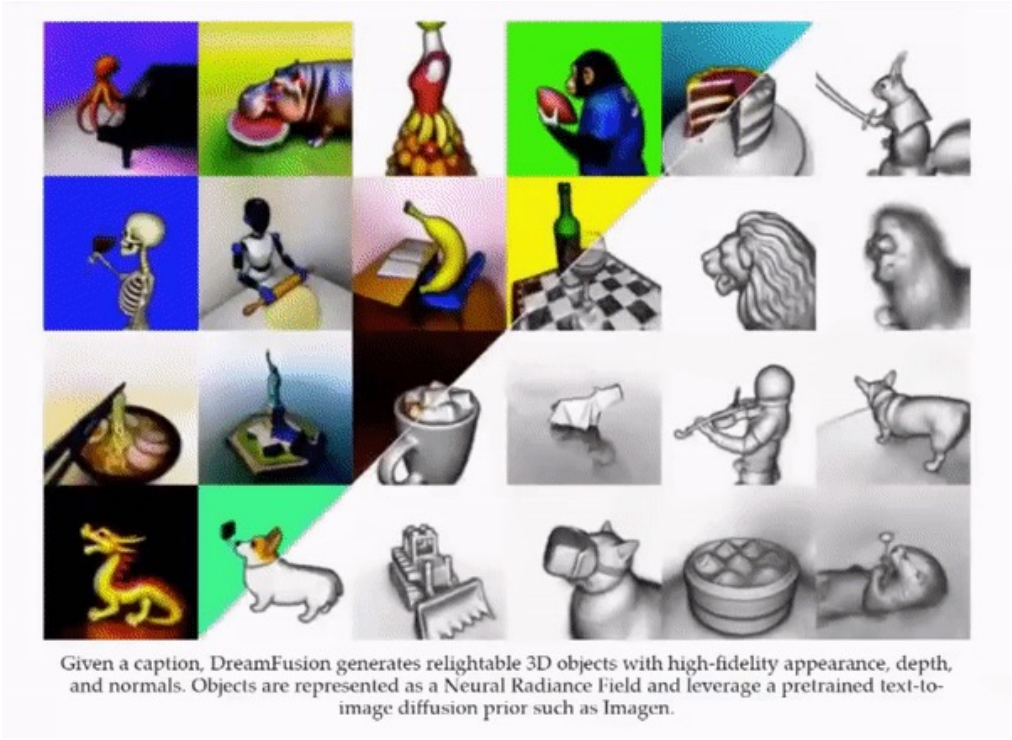
Prompt: the late afternoon sun peeking through the window of a New York City loft

W.A.L.T

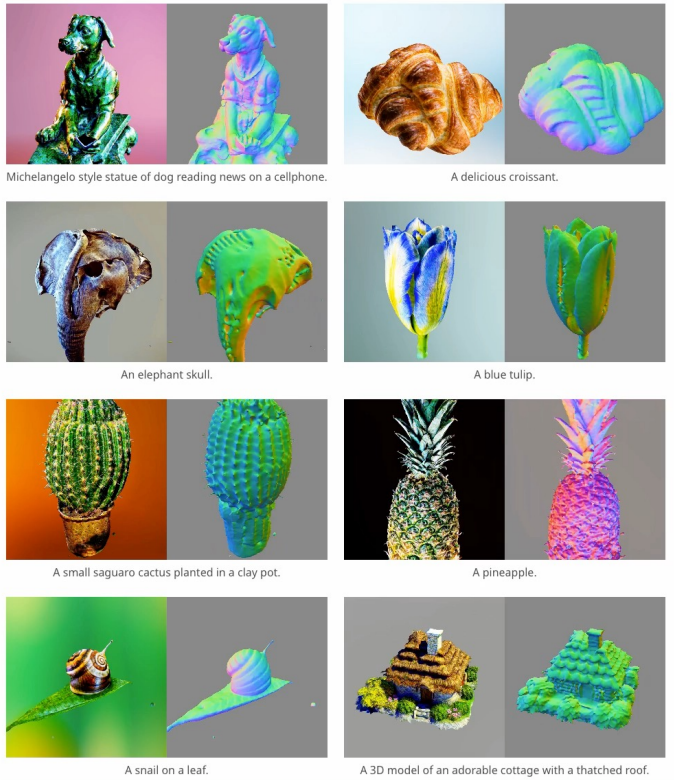
https://pika.art/; https://research.runwayml.com/gen2; https://walt-video-diffusion.github.io/

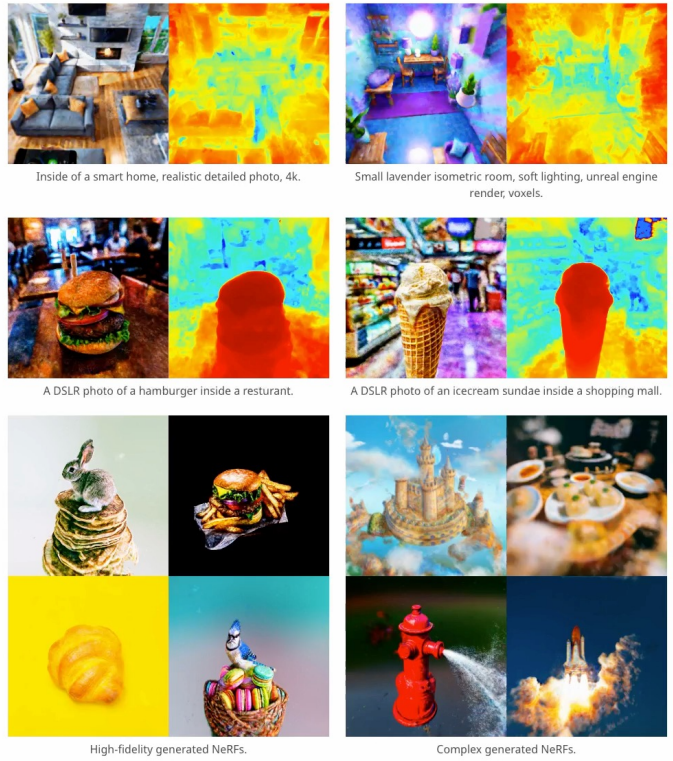
背景：生成式AI正迅速发展

3D内容生成



DreamFusion by Google





ProlificDreamer by Tsinghua

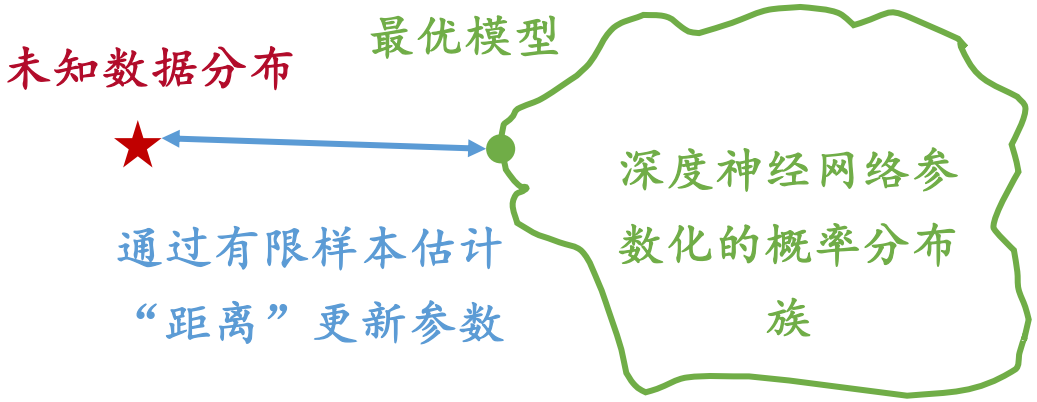
https://dreamfusion3d.github.io/index.html; https://ml.cs.tsinghua.edu.cn/prolificdreamer/

背景：生成式AI的内核

深度生成模型

- 核心问题：高维、复杂的联合概率分布的表示、学习与推断

经典图像数据：超过十万维的多峰分布



背景：深度生成模型的发展

Hidden Markov Models
Helmholtz Machine

VAE
(Variational Auto-encoders)

PixelCNN

Diffusion models;
Score-based generative models

1980s-2000s — 2013 — 2016 — 2020 — 2021-2023

2000s-2010s — 2014 — 2018 — 2020 — Text-to-image

Restricted Boltzmann Machine
Deep Belief Networks

GAN
(Generative Adversarial Networks)

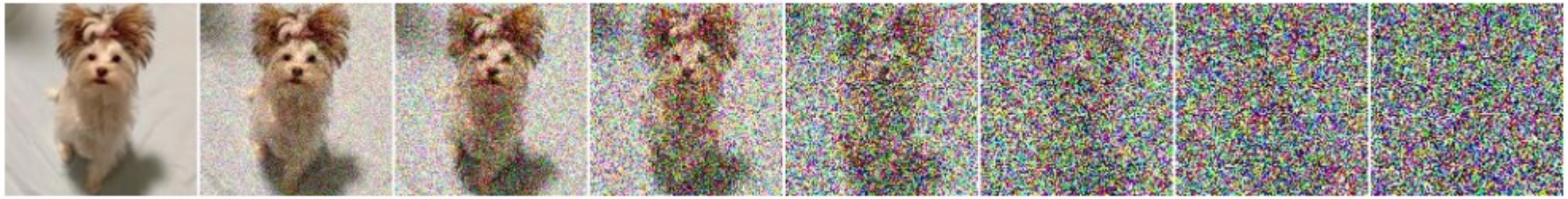
Normalizing Flows
(RealNVP; Glow)

Image-GPT

- DALL·E
- DALL·E-2
- Imagen
- Stable Diffusion

扩散模型的定义：加噪

前项链：高斯核马尔科夫链，一般 1000 步



数据分布

$q\left(\mathbf{x}^{(0)}\right)$

加入噪声

高斯噪声

$q\left(\mathbf{x}^{(T)}\right) \approx \mathcal{N}\left(\mathbf{x}^{(T)}; 0, \mathbf{I}\right)$

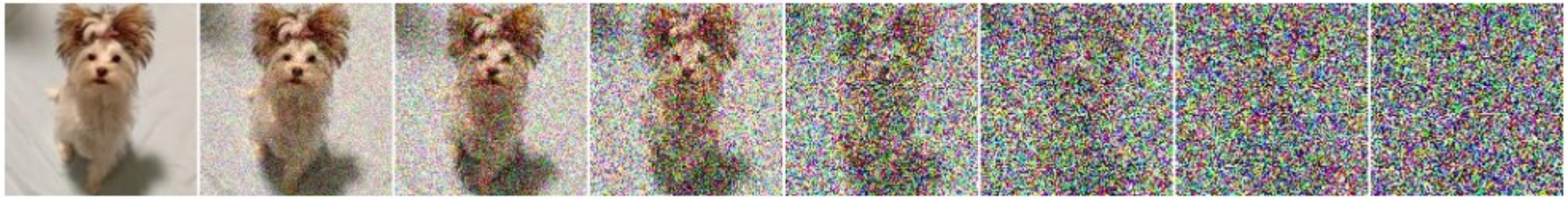
转移概率

$$q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right) = \mathcal{N}\left(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t\right)$$

Sohl-Dickstein et al., Deep unsupervised learning using nonequilibrium thermodynamics. ICML 2015
Ho et al. Denoising diffusion probabilistic models (DDPM), NeurIPS 2020
Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021

扩散模型的学习：去噪

可学习的反向高斯核马尔科夫链



模型分布

$$p\left(\mathbf{x}^{(0)}\right) \approx q\left(\mathbf{x}^{(0)}\right)$$

学习去噪

高斯先验分布

$$p\left(\mathbf{x}^{(T)}\right) = \mathcal{N}\left(\mathbf{x}^{(T)}; 0, \mathbf{I}\right)$$

参数化转移概率

$$p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right) = \mathcal{N}\left(\mathbf{x}^{(t-1)}; f_\mu\left(\mathbf{x}^{(t)}, t\right), f_\Sigma\left(\mathbf{x}^{(t)}, t\right)\right)$$

时间共享参数的神经网络作为高斯核，如何学习模型参数？

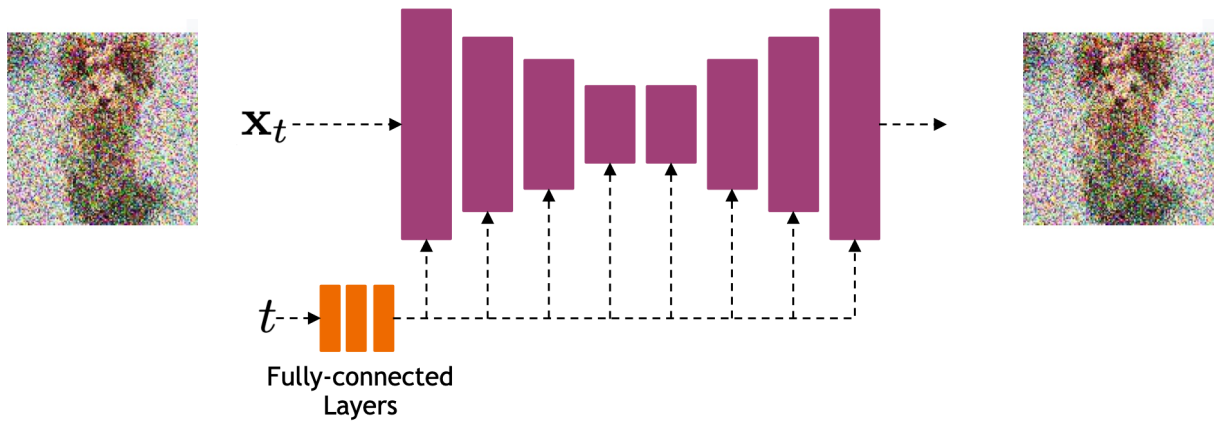
扩散模型的学习目标：最大似然估计

层次化隐变量模型（**hierarchical VAEs**），$p(x_0) = \int p(x_0|x_1)p(x_1|x_2)\dots p(x_T)dx_{1\dots T}$

=> 最大化变分下界：

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] \geq \mathrm{E}_{q(x_0)}\mathrm{E}_{q(x_{1:T}|x_0)} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \approx \mathrm{E}_{p_D(x_0),\epsilon}\, \mathrm{E}_{t\sim U[1,2,3\dots,T]} \|\mu_\theta(x_t,t) - \mu(x_t,x_0)\|^2$$

只预测均值，手工设置方差，例如对应前向分布噪声层级 $\beta_t \boldsymbol{I}$（对应特殊数据分布最优解）



$\mu_\theta(x_t, t)$

均值预测网络

Sohl-Dickstein et al., Deep unsupervised learning using nonequilibrium thermodynamics. ICML 2015

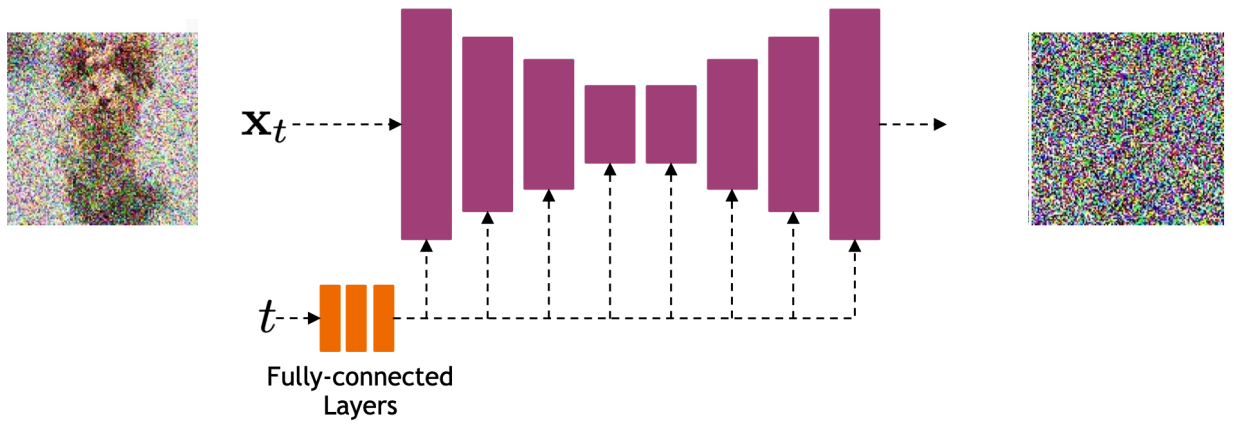
扩散模型的学习目标：噪声预测训练

均值回归等价于噪声回归：$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right)$

=>同时学习 **1000** 个噪声层级不同的去噪任务（可联系到**denoising score matching**）

$$\|\mu_\theta(x_t, t) - \mu(x_t, x_0)\|^2 \Leftrightarrow \|\epsilon_\theta(x_t, t) - \epsilon\|^2$$

噪声预测网络　高斯噪声



$\epsilon_\theta(x_t, t)$

噪声预测网络

Ho et al. Denoising diffusion probabilistic models (DDPM), NeurIPS 2020

扩散模型：算法

Algorithm 1 Training

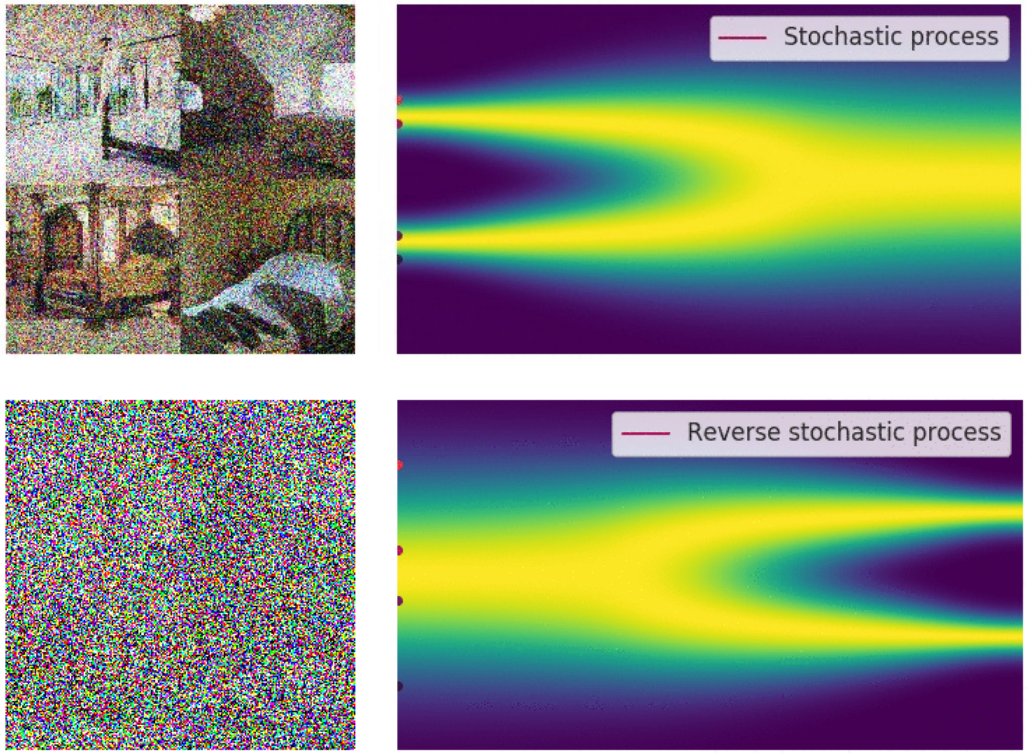
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0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

Algorithm 2 Sampling

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

Ho et al. Denoising diffusion probabilistic models (DDPM), NeurIPS 2020

连续时间扩散模型：无限步数



前向过程：

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

反向过程：

score function

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

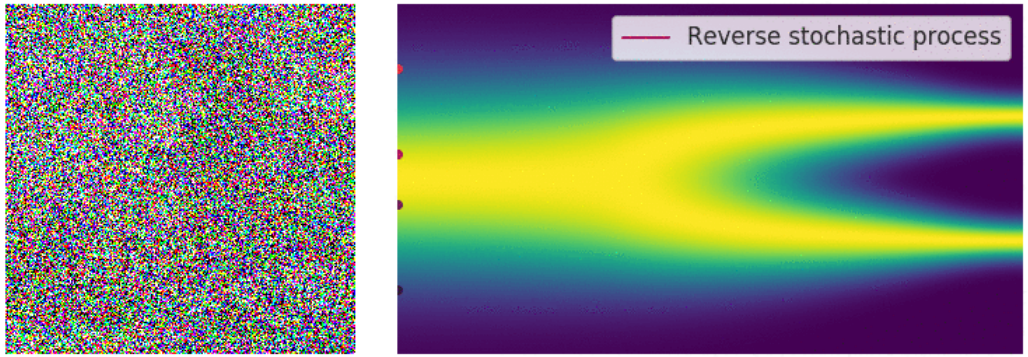
- 基于神经网络近似得分函数即可得到生成模型

训练目标：

$$\mathrm{E}_{p_D(x_0),\epsilon}\, \mathrm{E}_{t\sim U[1,2,3\ldots,T]} \|\epsilon_\theta(x_t, t) - \epsilon\|^2 \quad \xrightarrow{\text{无限步}} \quad \frac{1}{2}\int_0^T \omega(t)\mathbb{E}_{q_0(\boldsymbol{x}_0)}\mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}\|_2^2\right]\mathrm{d}t$$

Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021

连续时间扩散模型：无限步数



Diffusion SDE:

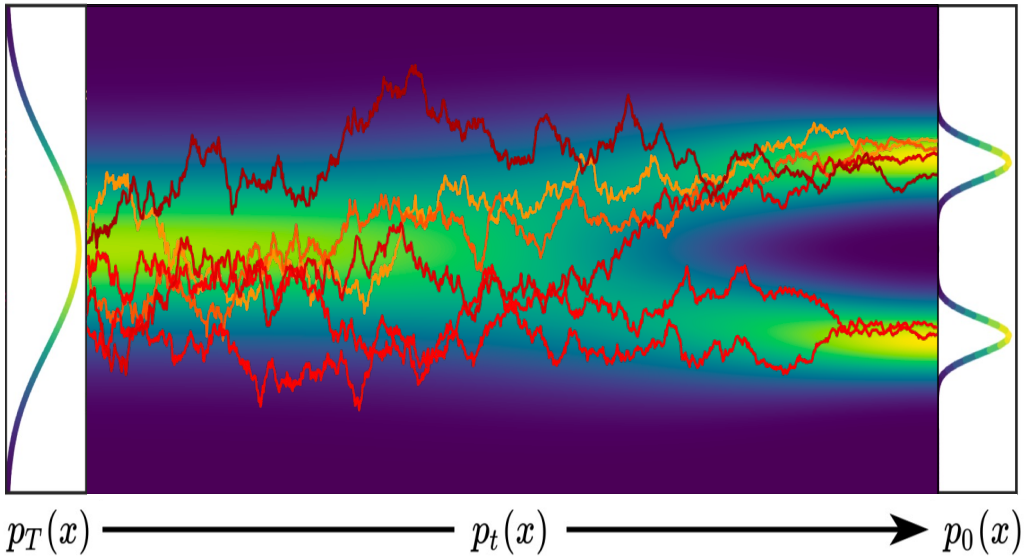
$$\mathrm{d}\boldsymbol{x}_t = \left[f(t)\boldsymbol{x}_t + \frac{g^2(t)}{\sigma_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}}_t, \quad \boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2\boldsymbol{I})$$

Diffusion ODE（Probability Flow ODE，确定性，边际分布一致）：

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = f(t)\boldsymbol{x}_t + \frac{g^2(t)}{2\sigma_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t), \quad \boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2\boldsymbol{I})$$

Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021

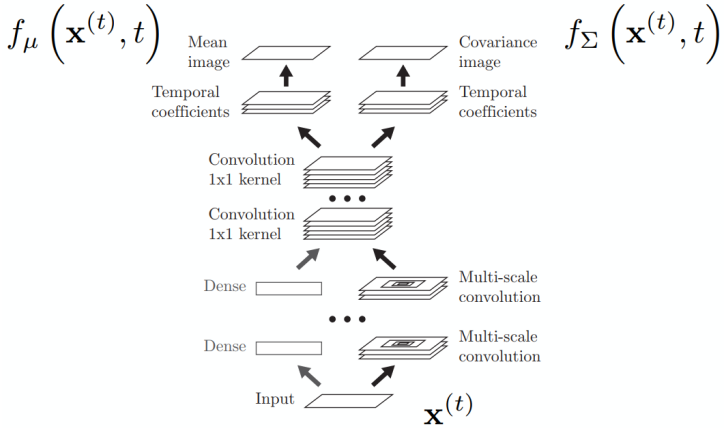
连续时间扩散模型：无限步数

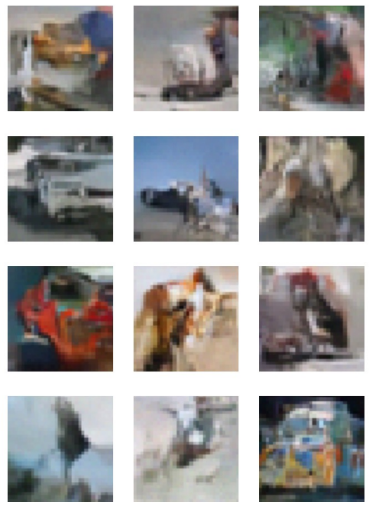


- **对上述SDE、ODE进行离散化求解即可实现采样**
- **SDE：**
  - 优点：持续的噪声注入可以帮助在扩散过程中弥补错误
  - 缺点：通常较慢，因为随机项在求解过程中需要细致的离散化
- **ODE：**
  - 优点：可以利用快速的ODE求解器，在需要非常快速的采样时效果最好
  - 缺点：没有"随机"错误校正，通常性能略低于随机采样

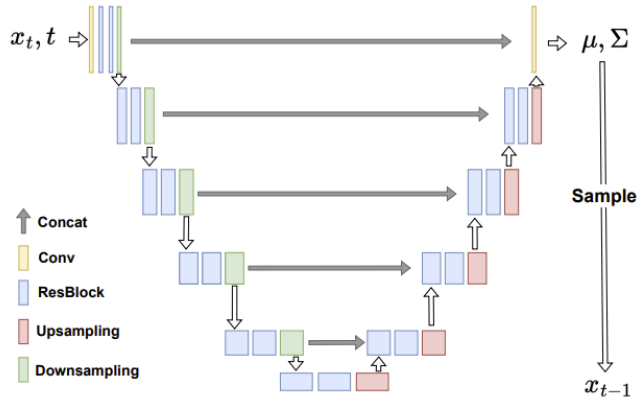
Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021

扩散模型的网络结构

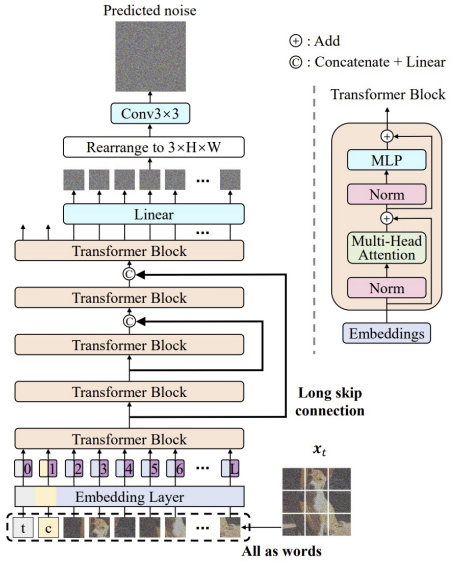




2015： 多尺度多通路的卷积网络



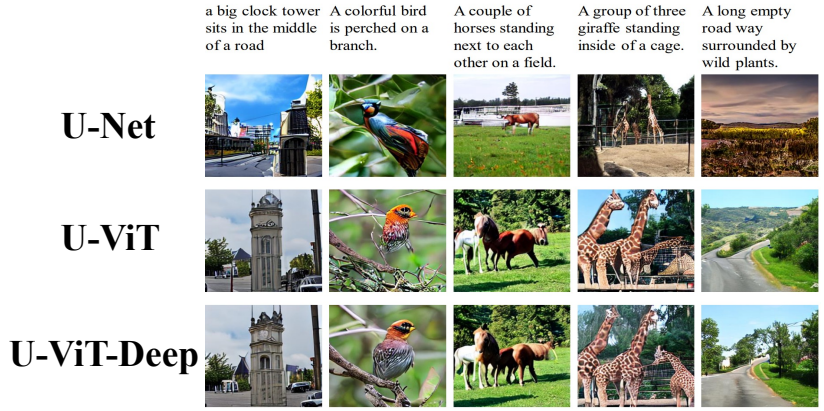
2021： U-Net变成主流



2023：U-ViT出现

简单、易扩展

- 方便大规模训练
- 方便跨模态训练



MS-COCO 文到图领先的生成结果

Bao F et al. All are Worth Words: A ViT Backbone for Diffusion Models. CVPR 2023.

扩散模型的优点总结

- 编码器$\boldsymbol{q(z|x)}$固定，不必学习
  - **VAEs 需要同时学习$\boldsymbol{q(z|x)}$和$\boldsymbol{p(x|z)}$**
- 训练目标足够简单
  - **MSE loss: $\|\epsilon_\theta(x_t, t) - \epsilon\|^2$**
- 收敛保证
  - 当总步数T足够多，反向过程的转移概率是高斯的
- 先进的网络结构
- 足够好的生成性能

扩散模型的方法研究：更可靠、普适、高效

- 基础改进（**30min**）
  - 对预训练分数预测模型的完善、校准、增强
  - 面向更高分辨率、多模态
  - …
- 生成加速（**30min**）
  - 设计**ODE Solver**
  - 模型蒸馏、一致性模型
  - …

一般扩散模型（如：**DDPM**）只学习逆向过程转移概率的高斯均值$\mu_\theta(x_t, t)$，而方差采用手工设计，这是次优的

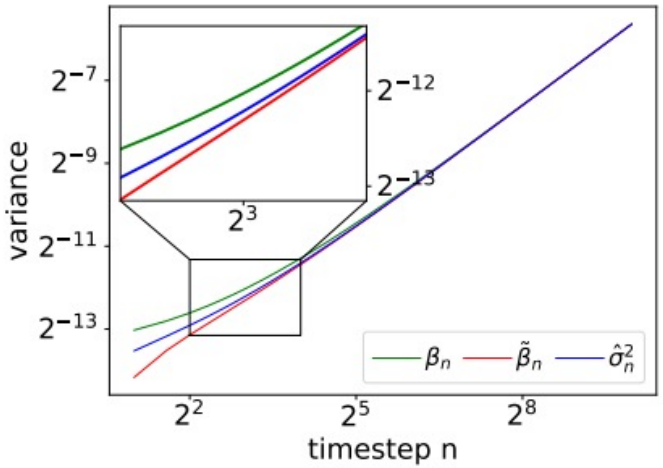
Analytic-DPM：扩散模型的最优采样方差理论（隐变量模型视角出发）

- 证明最大似然意义下最优采样方差闭式解，改变了手工设计方差的范式

定理：扩散概率模型在最大似然意义下关于评分函数/去噪函数的最优采样方差闭式解如下：

$$\sigma_t^{*2} = \frac{\beta_t}{1-\beta_t}\left(1 - \beta_t \mathrm{E}_{q_t(x_t)} \frac{\|\nabla \log q_t(x_t)\|^2}{d}\right).$$

基于预训练模型和蒙特卡洛方法近似

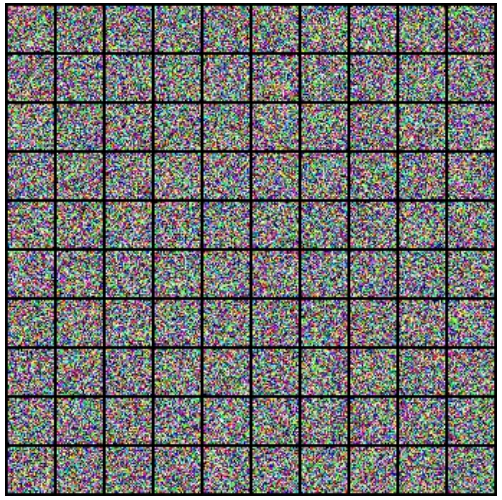


最优方差（蓝色）与手工方差在零时刻（即数据分布）附近有显著区别

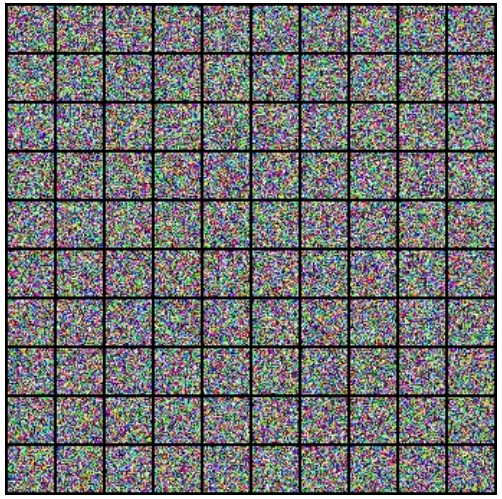
Bao F, Li C, Zhu J, et al. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. ICLR 2022.

Analytic-DPM：结果

- 无需额外训练，保证合成样本质量不变，相对**DDPM**至多加速 **20-80** 倍
- 作为核心技术部署于文到图生成大模型 **DALLE·2**



经典方法 **1000** 步



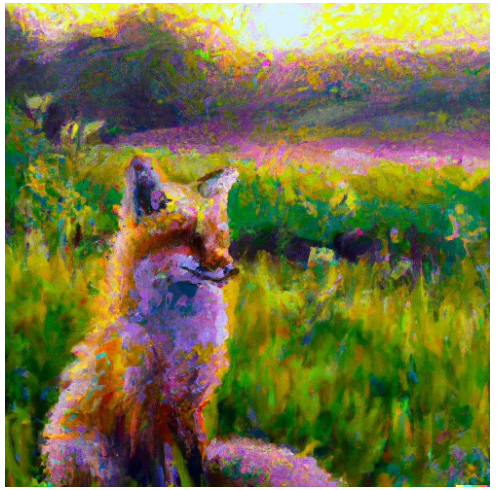
所提方法 **50** 步

To obtain a full generative model of images, we combine the CLIP image embedding *decoder* with a *prior* model, which generates possible CLIP image embeddings from a given text caption. We compare our text-to-image system with other systems such as DALL-E [40] and GLIDE [35], finding that our samples are comparable in quality to GLIDE, but with greater diversity in our generations. We also develop methods for training diffusion priors in latent space, and show that they achieve comparable performance to autoregressive priors, while being more compute-efficient. We refer to our full text-conditional image generation stack as *unCLIP*, since it generates images by inverting the CLIP image encoder.

For the AR prior, we use a Transformer text encoder with width 2048 and 24 blocks and a decoder with a causal attention mask, width 1664, and 24 blocks. For the diffusion prior, we use a Transformer with width 2048 and 24 blocks, and sample with Analytic DPM [2] with 64 strided sampling steps. To reuse hyperparameters tuned for diffusion noise schedules on images from Dhariwal and Nichol [11], we scale the CLIP embedding inputs by 17.2 to match the empirical variance of RGB pixel values of ImageNet images scaled to $[-1, 1]$.

	AR prior	Diffusion prior	64	$64 \rightarrow 256$	$256 \rightarrow 1024$
Diffusion steps	-	1000	1000	1000	1000
Noise schedule	-	cosine	cosine	cosine	linear
Sampling steps	-	64	250	27	15
Sampling variance method	-	analytic [2]	learned [34]	DDIM [47]	DDIM [47]
Crop fraction	-	-	-	0.25	0.25
Model size	1B	1B	3.5B	700M	300M
Channels	-	-	512	320	192
Depth	-	-	3	3	2
Channels multiple	-	-	1,2,3,4	1,2,3,4	1,1,2,2,4,4
Heads channels	-	-	64	-	-
Attention resolution	-	-	32,16,8	-	-
Text encoder context	256	256	256	-	-
Text encoder width	2048	2048	2048	-	-
Text encoder depth	24	24	24	-	-
Text encoder heads	32	32	32	-	-

显著加速 **DALLE·2**



“a painting of a fox sitting in a field at sunrise in the style of Claude Monet”

在扩散模型的实际训练中，使用$\boldsymbol{q_0(x_0)}$采样的数据集近似损失函数$\frac{1}{2}\int_0^T \omega(t)\mathbb{E}_{q_0(\boldsymbol{x}_0)}\mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{\epsilon}\|_2^2\right]\mathrm{d}t$中的期望，为训练模型引入偏差

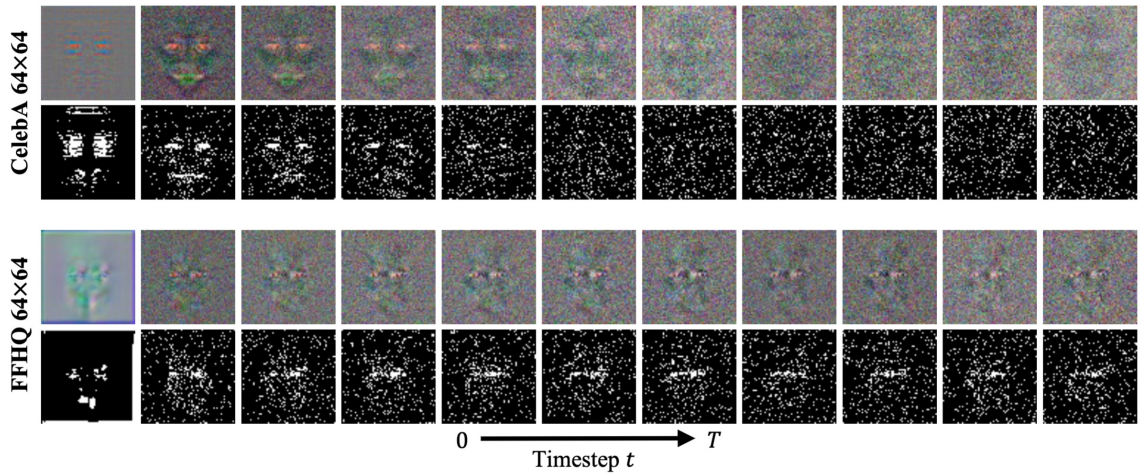
Calibrated-DPMs：扩散模型的校准（偏差消除）

观察：在温和的边界条件假设下，有

$$\mathrm{E}_{q_0(x_0)}\nabla_{x_0}\log q_0(x_0) = 0.$$

定理：$\alpha_t\nabla_{x_t}\log q_t(x_t)$的随机过程是关于$x_t$的逆时间过程的鞅，则

$$\mathrm{E}_{q_t(x_t)}\nabla_{x_t}\log q_t(x_t) = 0.$$



然而，实际训练的分数预测模型并非具有0均值

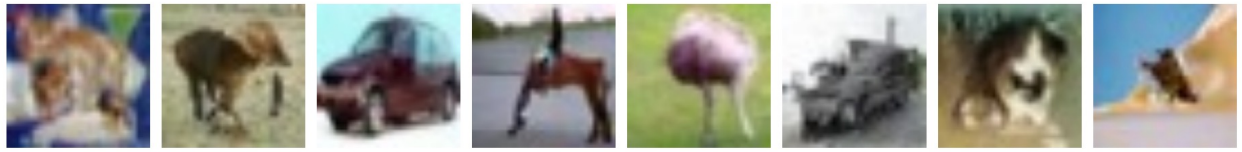
- 将训练完的模型$\epsilon_\theta(x_t,t)$校准为$\epsilon_\theta(x_t,t) - \mathrm{E}_{q_t(x_t)}\epsilon_\theta(x_t,t)$，可降低模型的**score matching loss**，改善模型似然

基于蒙特卡洛方法或额外训练一个模型近似

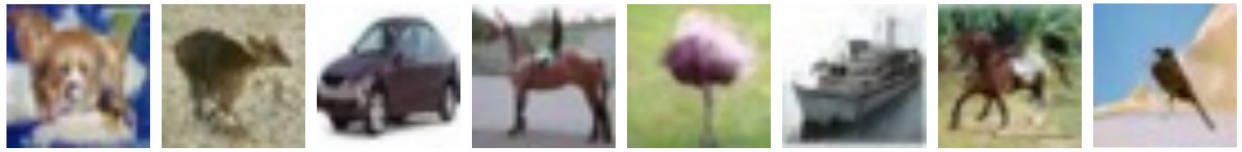
Pang, Lu, Du, Lin, Yan, Deng. On Calibrating Diffusion Probabilistic Models. NeurIPS 2023.

Calibrated-DPMs：结果

- 无需额外训练，在多个数据集上提高合成样本质量，相对未校准模型提高 **FID至多1.1**



w/o calibration (**baseline**)



w/ calibration (**ours**)

Table 1: Comparison on sample quality measured by FID ↓ with varying NFE on CIFAR-10. Experiments are conducted using a linear noise schedule on the discrete-time model from [15]. We consider three variants of DPM-Solver with different orders. The results with † mean the actual NFE is order $\times \lfloor \frac{\text{NFE}}{\text{order}} \rfloor$ which is smaller than the given NFE, following the setting in [26].

Noise prediction	DPM-Solver	Number of evaluations (NFE) 10	15	20	25	30	35	40
$\epsilon_\theta^t(x_t)$	1-order	20.49	12.47	9.72	7.89	6.84	6.22	5.75
	2-order	7.35	†4.52	4.14	†3.92	3.74	†3.71	3.68
	3-order	†23.96	4.61	†3.89	†3.73	3.65	†3.65	†3.60
$\epsilon_\theta^t(x_t) - \mathbb{E}_{q_t(x_t)}[\epsilon_\theta^t(x_t)]$	1-order	19.31	11.77	8.86	7.35	6.28	5.76	5.36
	2-order	**6.76**	†4.36	4.03	†3.66	3.54	†3.44	3.48
	3-order	†53.50	**4.22**	**†3.32**	**†3.33**	**3.35**	**†3.32**	**†3.31**

Table 2: Comparison on sample quality measured by FID ↓ with varying NFE on CelebA 64×64. Experiments are conducted using a linear noise schedule on the discrete-time model from [35]. The settings of DPM-Solver are the same as on CIFAR-10.

Noise prediction	DPM-Solver	Number of evaluations (NFE) 10	15	20	25	30	35	40
$\epsilon_\theta^t(x_t)$	1-order	16.74	11.85	7.93	6.67	5.90	5.38	5.01
	2-order	**4.32**	†3.98	2.94	†2.88	2.88	†2.88	2.84
	3-order	†11.92	3.91	†2.84	†2.76	2.82	†2.81	†2.85
$\epsilon_\theta^t(x_t) - \mathbb{E}_{q_t(x_t)}[\epsilon_\theta^t(x_t)]$	1-order	16.13	11.29	7.09	6.06	5.28	4.87	4.39
	2-order	4.42	†3.94	2.61	†2.66	2.54	†2.52	**2.49**
	3-order	†35.47	**3.62**	**†2.33**	**†2.43**	**2.40**	**†2.43**	**†2.49**

低质量的生成广泛存在，现有指标衡量分布的质量（如：**IS**、**FID**）或不能可靠反映出真实视觉质量（如：**CLIPScore**、**ImageReward**）

准确高效地检测出低质量生成，对提升用户体验非常重要

Stable Diffusion v1.5
***ImageReward*: 1.74**

Preferred by human

Stable Diffusion XL Base
***ImageReward*: 0.70**

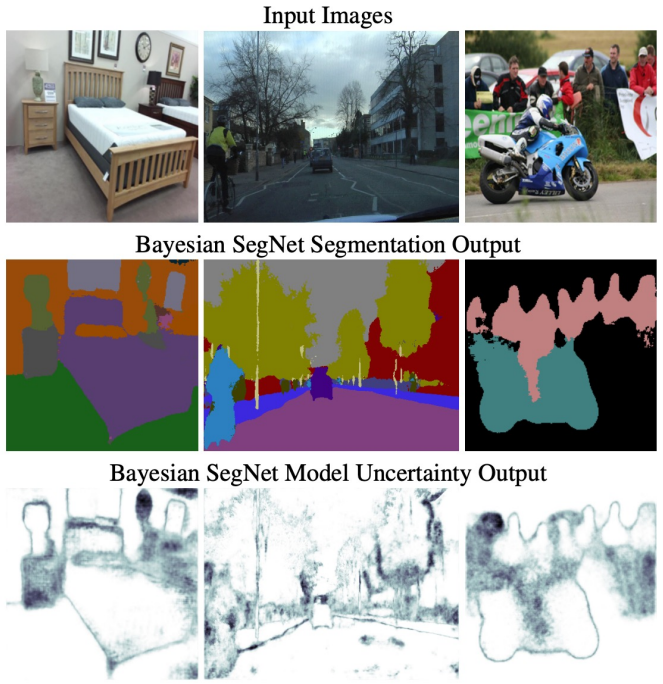
Stable Diffusion XL Base + Refiner
***ImageReward*: 0.75**

Prompt: A beautiful girl near the lake

低质量生成检测：从贝叶斯不确定性的视角

观察：

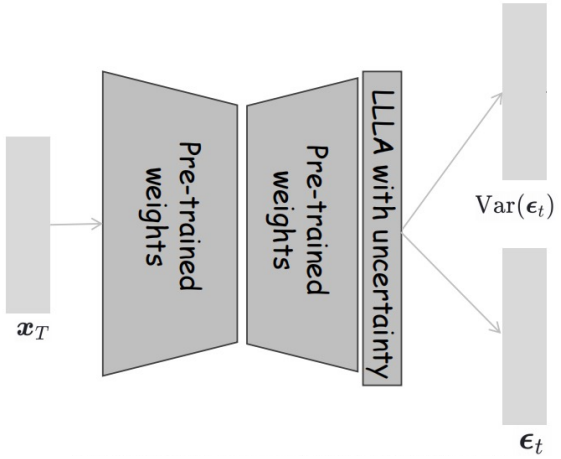
- 低质量的生成样本处于真实训练样本的流型之外，是分布外样本
- 基于贝叶斯不确定性可以有效检测分布外样本
- 贝叶斯深度学习方法在传统图到图变换的应用中可有效衡量不确定性



[Bayesian SegNet, Kendall et al.]

如何将贝叶斯深度学习技术融入扩散模型，实现基于贝叶斯不确定性的低质量生成检测？

基于贝叶斯推断量化扩散模型生成样本的不确定性



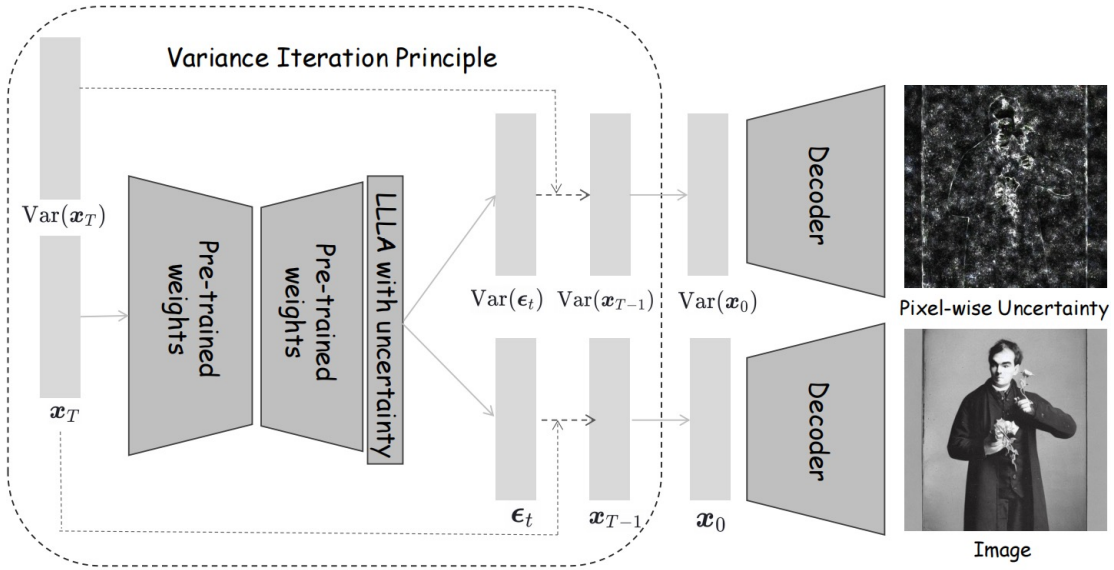
- 基于最后层拉普拉斯近似方法，对预训练得分预测模型进行低成本处理，将原始的点预测转化为高斯分布预测

扩散模型的推理不是一蹴而就的，需要迭代得到最终预测（如： DDIM ）：

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha'_t}}(\mathbf{x}_t - \frac{1-\alpha'_t}{\sqrt{1-\bar{\alpha}'_t}}\boldsymbol{\epsilon}_t) + \sqrt{\beta_t}\mathbf{z}$$

那么，如何量化最终生成内容的不确定性？

BayesDiff：扩散模型中预测不确定性的动力学刻画



- 高斯设定下：不确定性=方差
- 方差迭代：

$$\mathrm{Var}(\mathbf{x}_{t-1}) = \frac{1}{\alpha'_t}\mathrm{Var}(\mathbf{x}_t) - 2\frac{1-\alpha'_t}{\alpha'_t\sqrt{1-\bar{\alpha}'_t}}\mathrm{Cov}(\boldsymbol{x}_t, \boldsymbol{\epsilon}_t) + \frac{(1-\alpha'_t)^2}{\alpha'_t(1-\bar{\alpha}'_t)}\mathrm{Var}(\boldsymbol{\epsilon}_t) + \beta_t$$

$$\mathbb{E}(\boldsymbol{x}_{t-1}) = \frac{1}{\sqrt{\alpha'_t}}\mathbb{E}(\boldsymbol{x}_t) - \frac{1-\alpha'_t}{\sqrt{\alpha'_t(1-\bar{\alpha}'_t)}}\mathbb{E}(\boldsymbol{\epsilon}_t)$$

- 估计Cov项：拆分（全期望公式）+蒙特卡洛
- 跳跃机制（只在部分推理步骤进行方差迭代）
  - 在引入不超过**2x**推理开销的条件下，实现可靠的不确定性量化

Kou, Gan, Wang, Li, Deng. BayesDiff: Estimating Pixel-wise Uncertainty in Diffusion via Bayesian Inference. 2023

BayesDiff：结果

Figure 3: The images with the highest (left) and lowest (right) uncertainty among 5000 unconditional generations of U-ViT model trained on ImageNet at 256 × 256 resolution.

Figure 4: The images with the highest (left) and lowest (right) uncertainty among 80 generations on Stable Diffusion at 512 × 512 resolution.

像素上的不确定性相加得到图像不确定性，可有效检出低质量生成（如：复杂背景、模糊、歧义）

Table 1: Comparison on three metrics between randomly selected images and our selected images. We use 50 NFE for both DDIM and DPM-Solver sampler.

Model	Dataset	Sampler	FID ↓		Precision ↑		Recall ↑	
			random	ours	random	ours	random	ours
ADM	ImageNet 128	DDIM	8.65	8.48	0.661	0.665	0.655	0.653
ADM	ImageNet 128	2-order DPM-Solver	9.72	9.67	0.657	0.659	0.649	0.649
U-ViT	ImageNet 256	2-order DPM-Solver	7.21	6.81	0.698	0.705	0.658	0.657
U-ViT	ImageNet 512	2-order DPM-Solver	17.75	16.87	0.728	0.732	0.602	0.604

筛选后至多降低~0.9FID

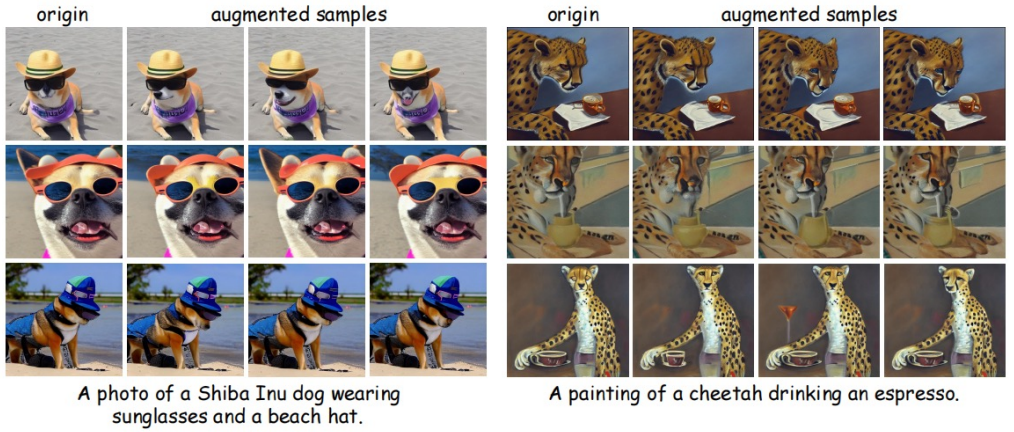
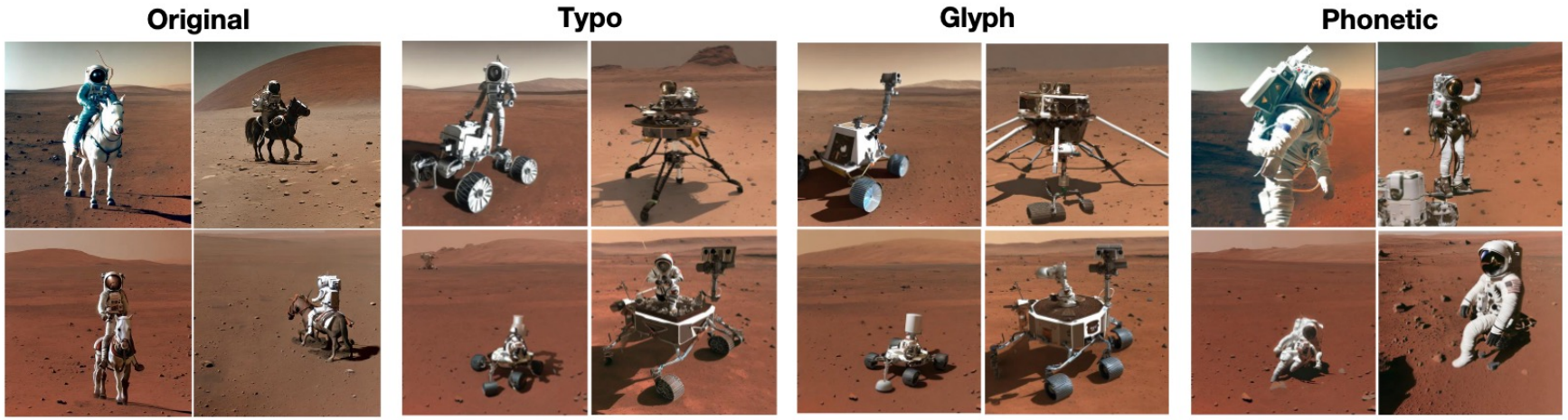


Figure 7: Examples of the augmentation of good generations with enhanced diversity on Stable Diffusion with DDIM sampler (50 NFE).

可以从中间的任意时刻隐状态的分布$\mathcal{N}(\mathbb{E}(\boldsymbol{x}_t), \mathrm{Var}(\boldsymbol{x}_t))$出发，进行重采样，实现低成本数据增广

扩散模型生成对输入提示的敏感程度，也
是影响用户体验的重要因素

文本-图像扩散模型的输入鲁棒性



Rule	Ori. Sentence	Adv. Sentence
Typo	A red ball on green grass under a blue sky.	A rde ball on green grass under a blue skky.
Glyph	A red ball on green grass under a blue sky.	A rêd ball 0n green grass under a blue sky.
Phonetic	A red ball on green grass under a blue sky.	A read ball on green grass under a blue SKY.

基于typo、glyph、phonetic来模仿现实世界扰动

Gao et al., Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks, 2023.

文本-图像扩散模型的输入鲁棒性

- 优化目标
  - **MMD距离:**

$$D_{MMD}(p(x|c')\|p(x|c)) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i, x_j) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i, \bar{x}_j)$$

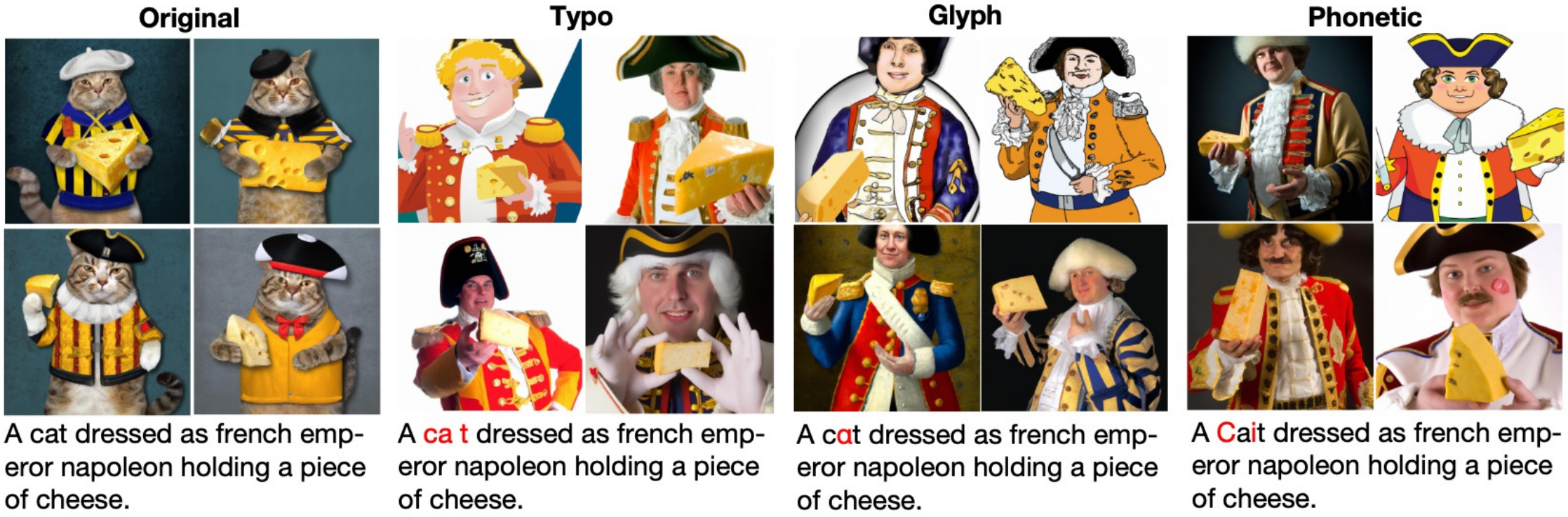
  - **KL:**

$$D_{KL}(p(x|c')\|p(x|c)) \approx \alpha\left[\frac{1}{N}\sum_{1}^{N}h_\phi(x_i)\right]^{\top}\left(g_\phi(c') - g_\phi(c)\right)$$

  - **Two-sample test:** $\hat{t}(\{\varphi(x_i)\}_{i=1}^{N}, \{\varphi(\bar{x}_i)\}_{i=1}^{N})$
- 攻击方法
  - 词重要性排序
  - 词级别扰动

Gao et al., Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks, 2023.

文本-图像扩散模型的输入鲁棒性：结果



Real-world attacks on DALLE-2.

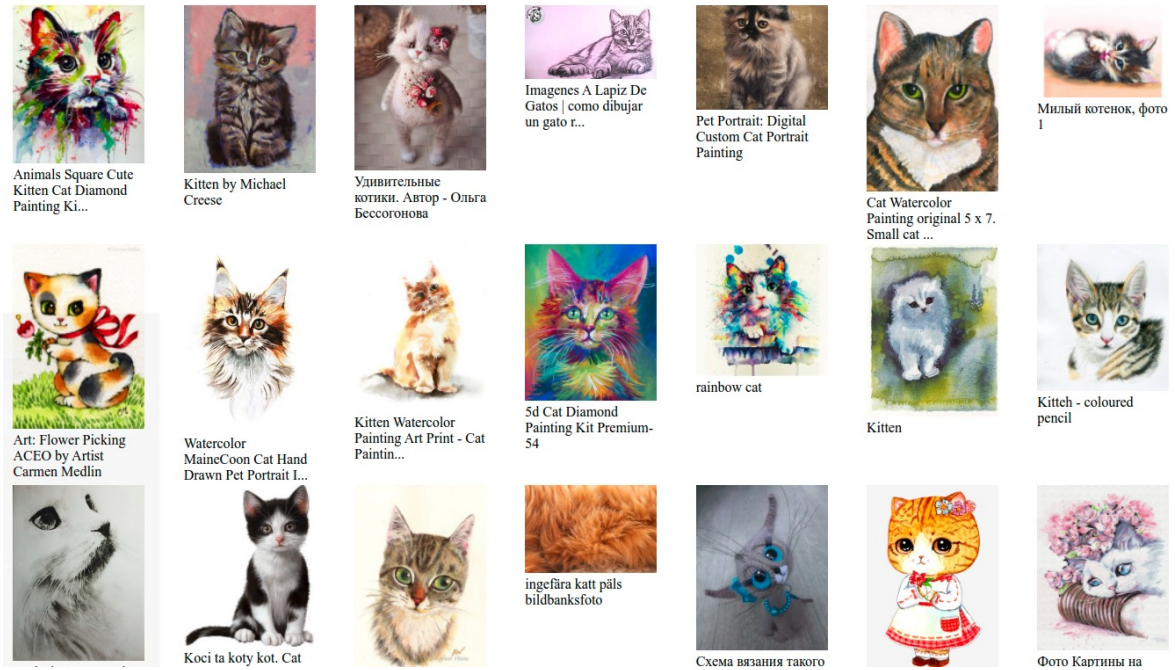
Gao et al., Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks, 2023.

文本-图像扩散模型的输入鲁棒性：结果

Human (random) attacks on Stable Diffusion.

Gao et al., Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks, 2023.

更高的分辨率和跨模态任务（如：文本到图像）
为扩散模型的模型构建和训练方法提出新的要求

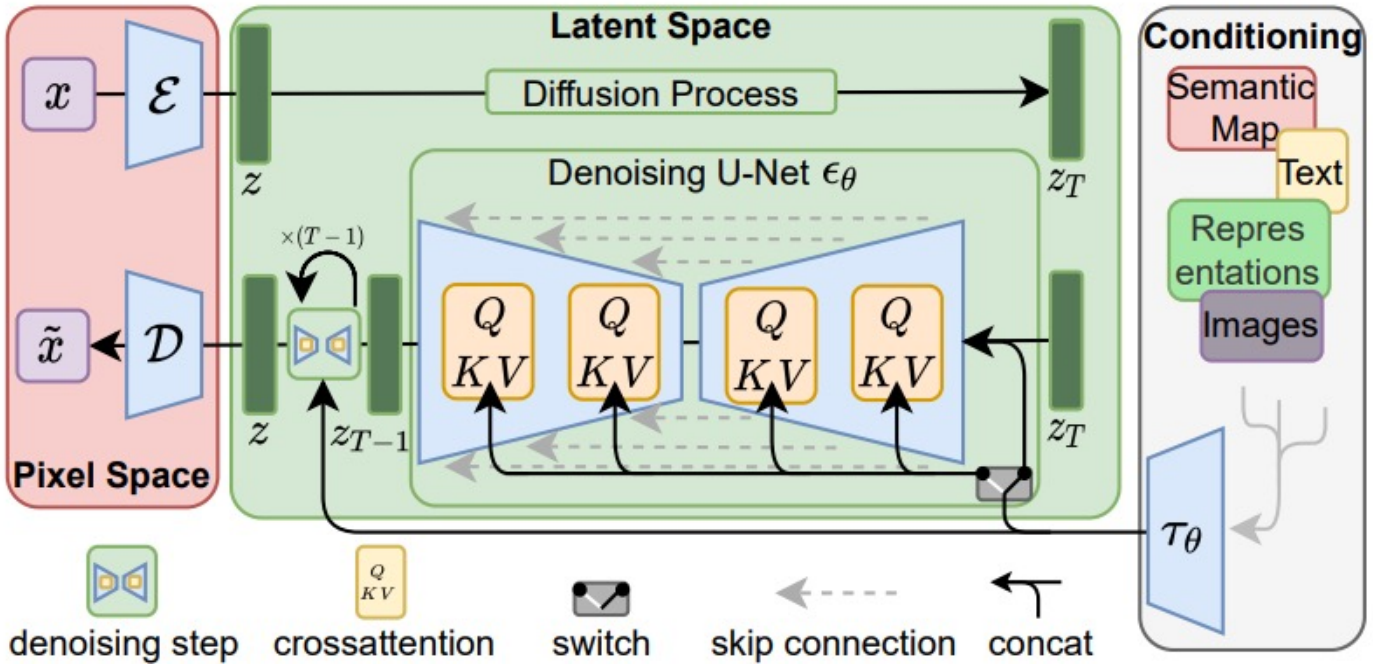


LAION：开放域、高噪声、大规模

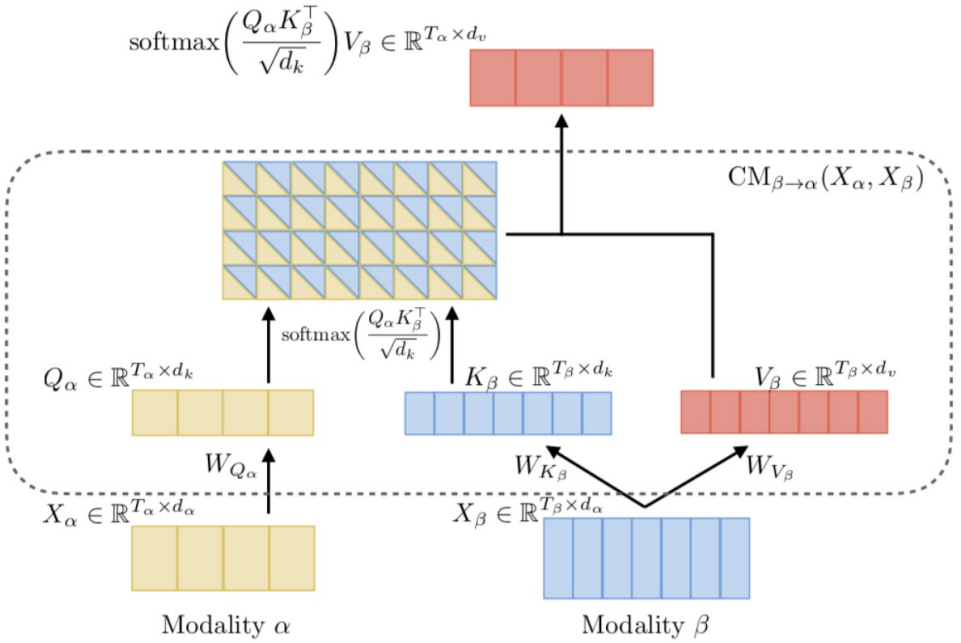
隐空间扩散模型

基于classifier-free guidance激发辅助信息：

$$\hat{\epsilon}_\theta(\mathbf{x}_t|y) = s \cdot \epsilon_\theta(\mathbf{x}_t|y) + (1-s) \cdot \epsilon_\theta(\mathbf{x}_t|0)$$
$$= \epsilon_\theta(\mathbf{x}_t|0) + s \cdot (\epsilon_\theta(\mathbf{x}_t|y) - \epsilon_\theta(\mathbf{x}_t|0))$$



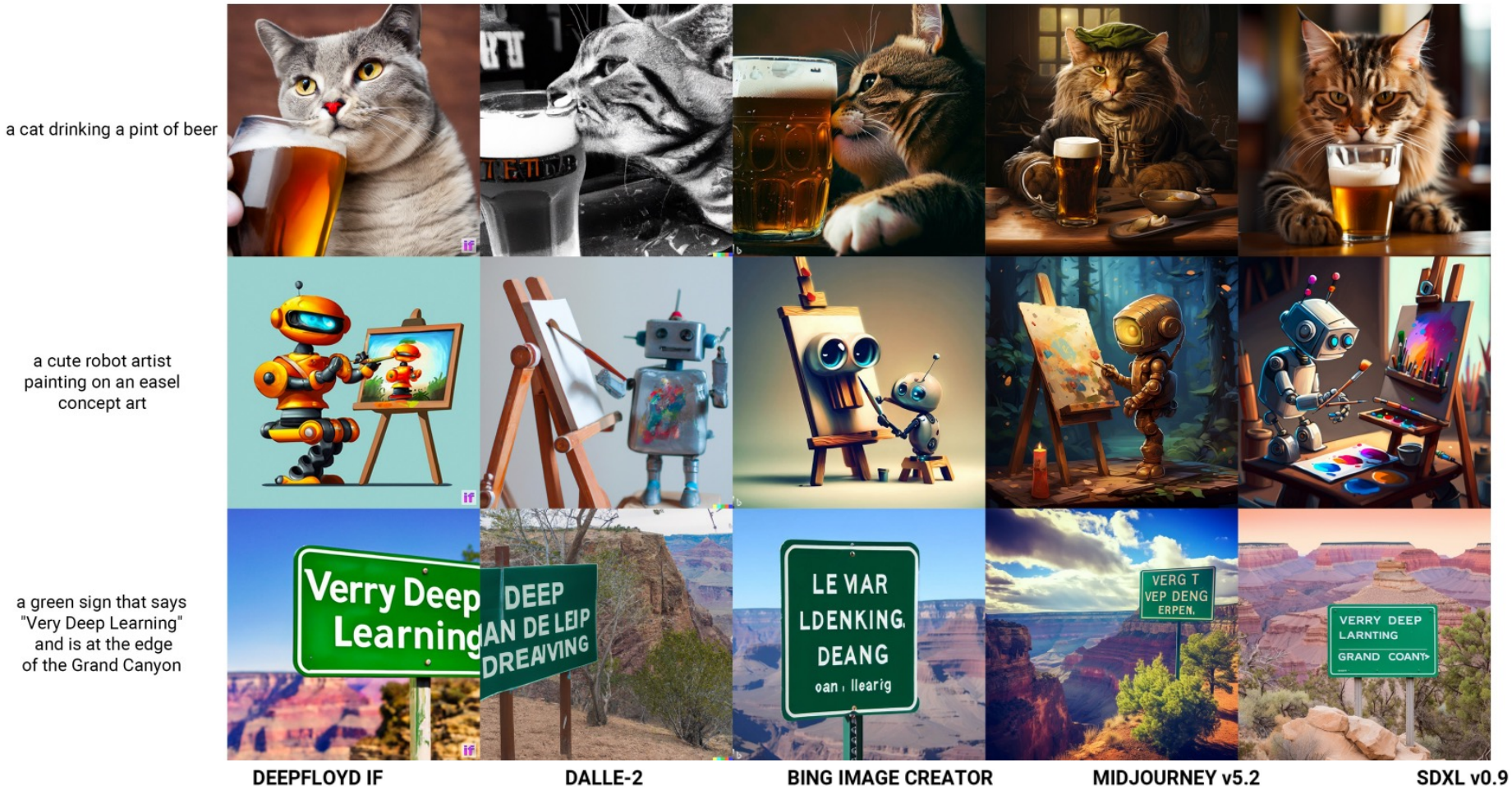
隐空间条件模型



Cross Attention 模块

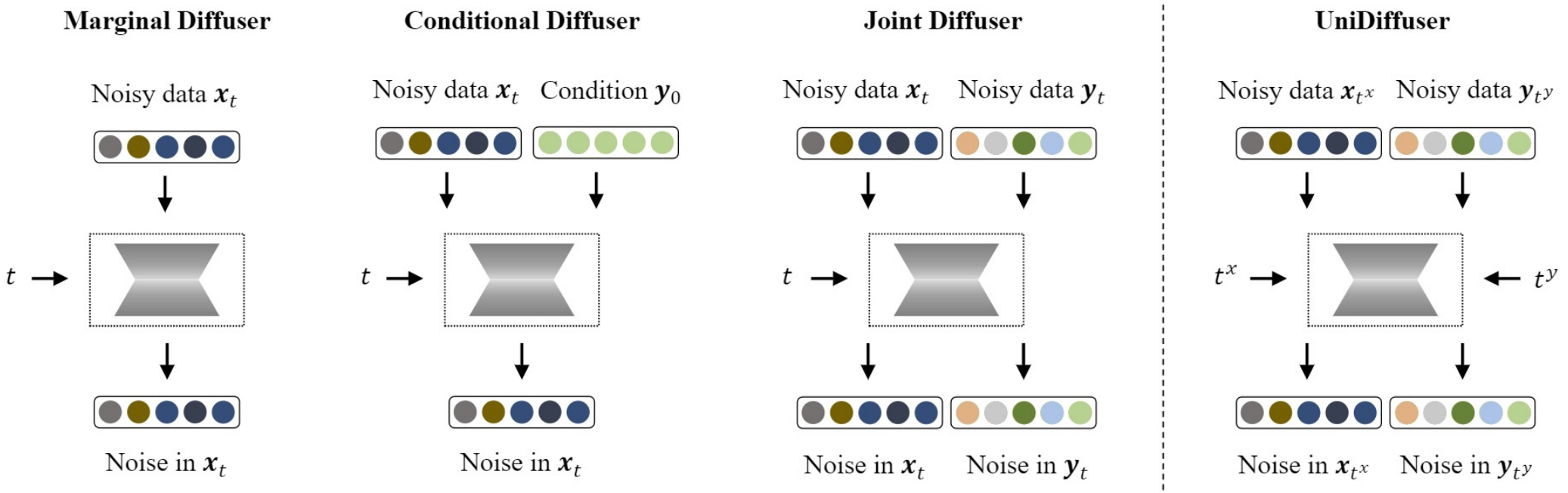
Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Mode. CVPR 2022.

SDXL 生成结果



大模型 + 大数据 + 适当算法/结构 = 强泛化、任务通用智能

多模态扩散概率模型：通用训练算法

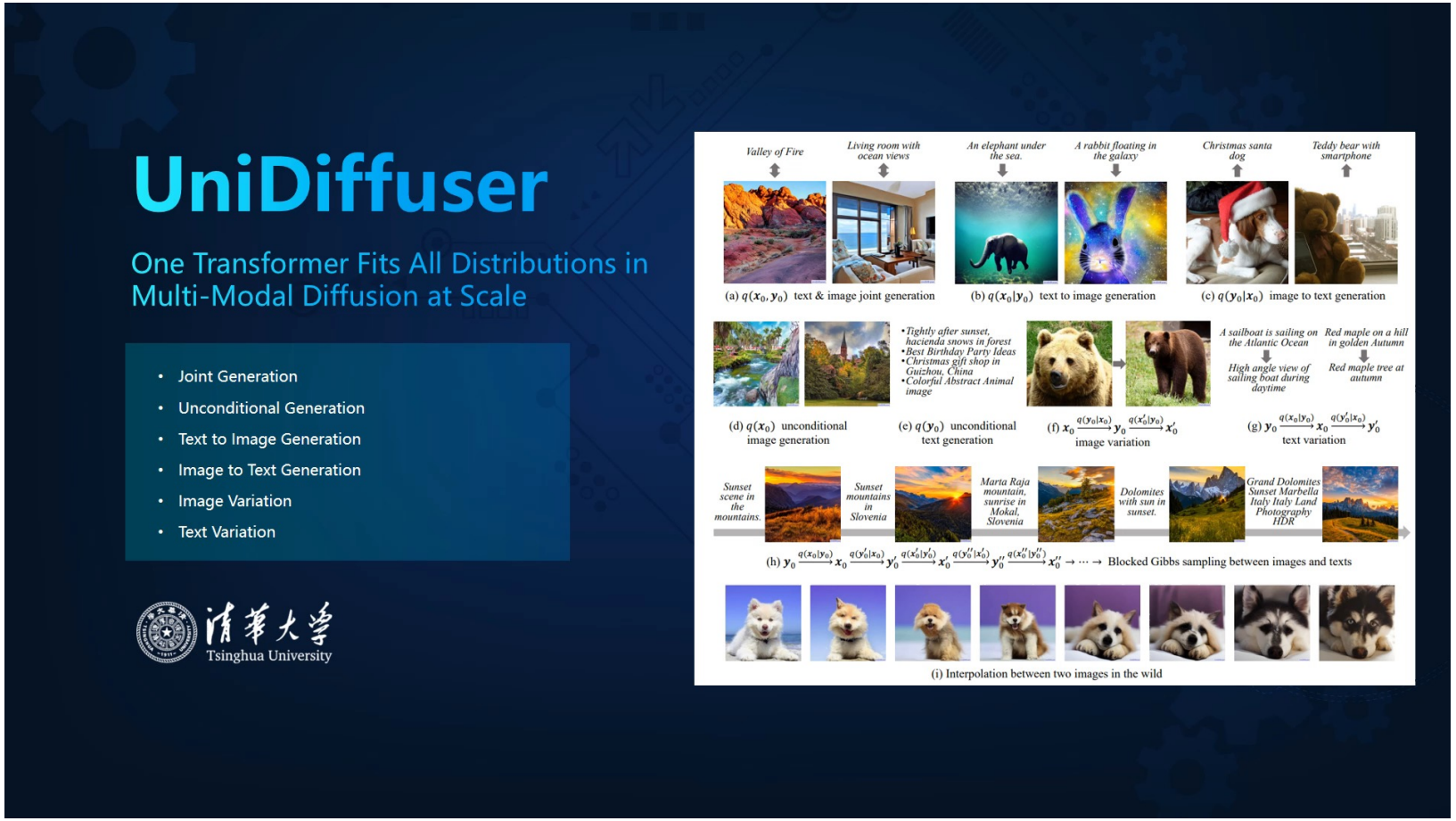


文本/图像生成　　文到图/图到文生成　　图文联合生成　　统一建模框架

Bao F et al. One transformer (U-ViT) fits all distributions in multi-modal diffusion. ICML 2023.

Unidiffuer：文本-图像通用扩散概率大模型

- 文图通用模型
  - 适当增大模型
  - 训练时间不变
  - 推断时间不变
  - 推断效果可比
  - 处理 5 种任务



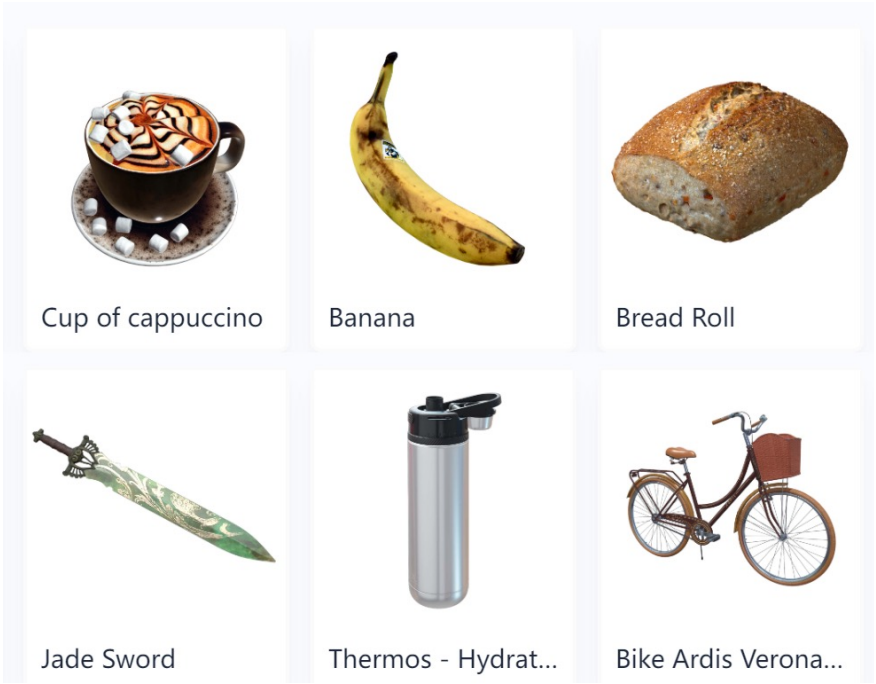
模态上的新挑战I：开放域文本到三维内容生成

A blue jay standing on a large basket of rainbow macarons

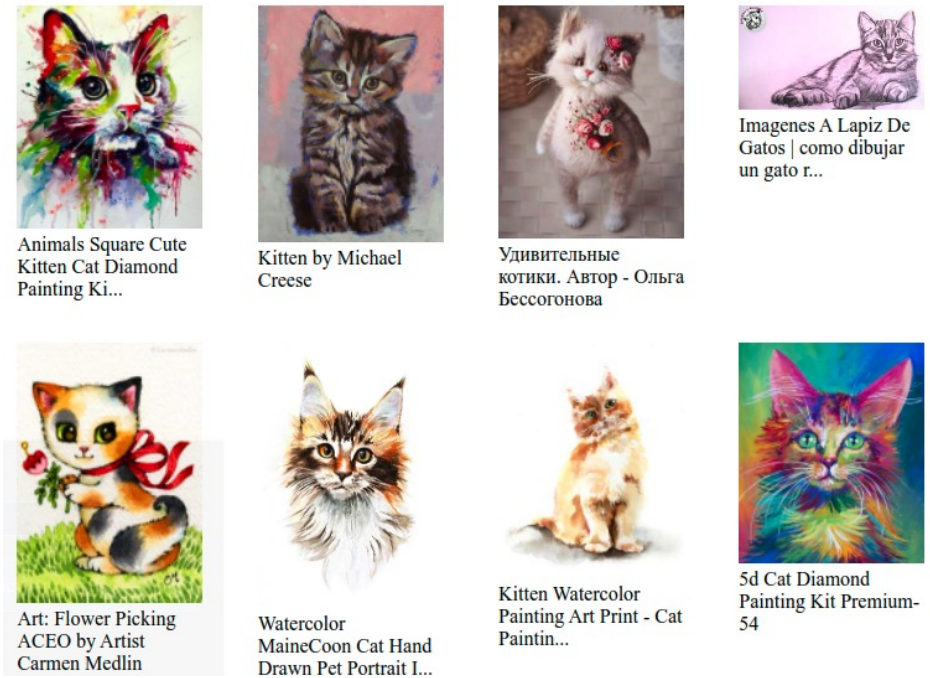
一只冠蓝鸦站在一大篮彩虹马卡龙上

挑战：数据量小

Objaverse：特定域、低噪声、小规模 (800K)



LAION：开放域、高噪声、大规模 (5B, 400M)

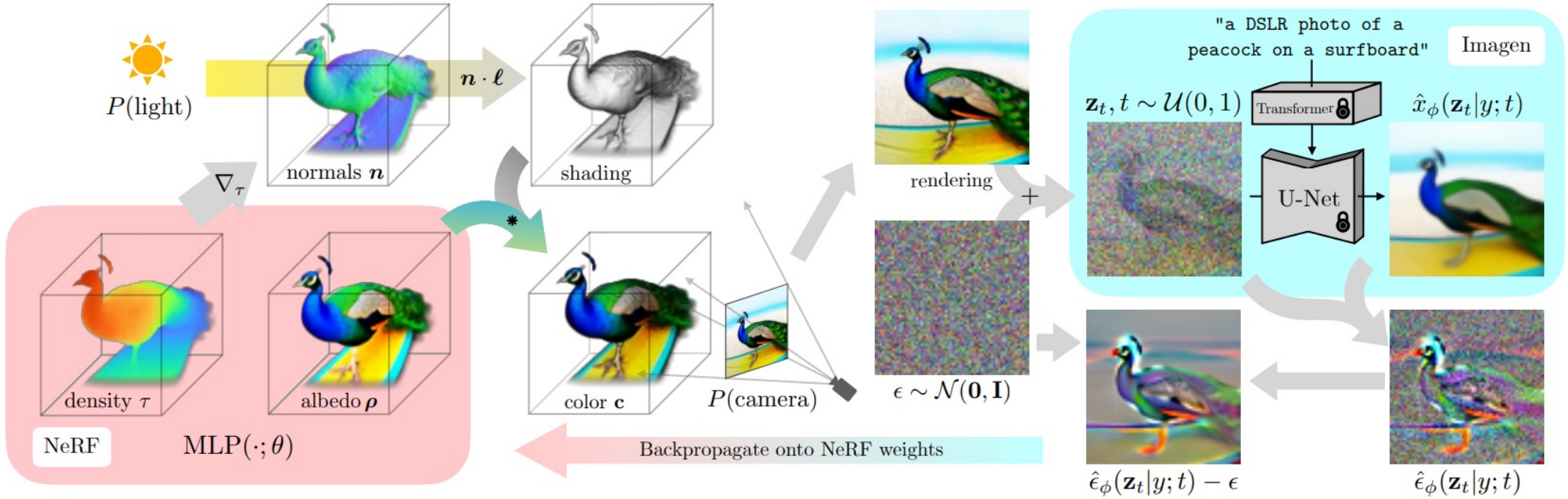


如何在样本量不充分的情况下泛化到开放域文本？

DreamFusion：将图像模型蒸馏为3D模型

三维结构表示与可微渲染过程

基础扩散模型定义的图像先验分布



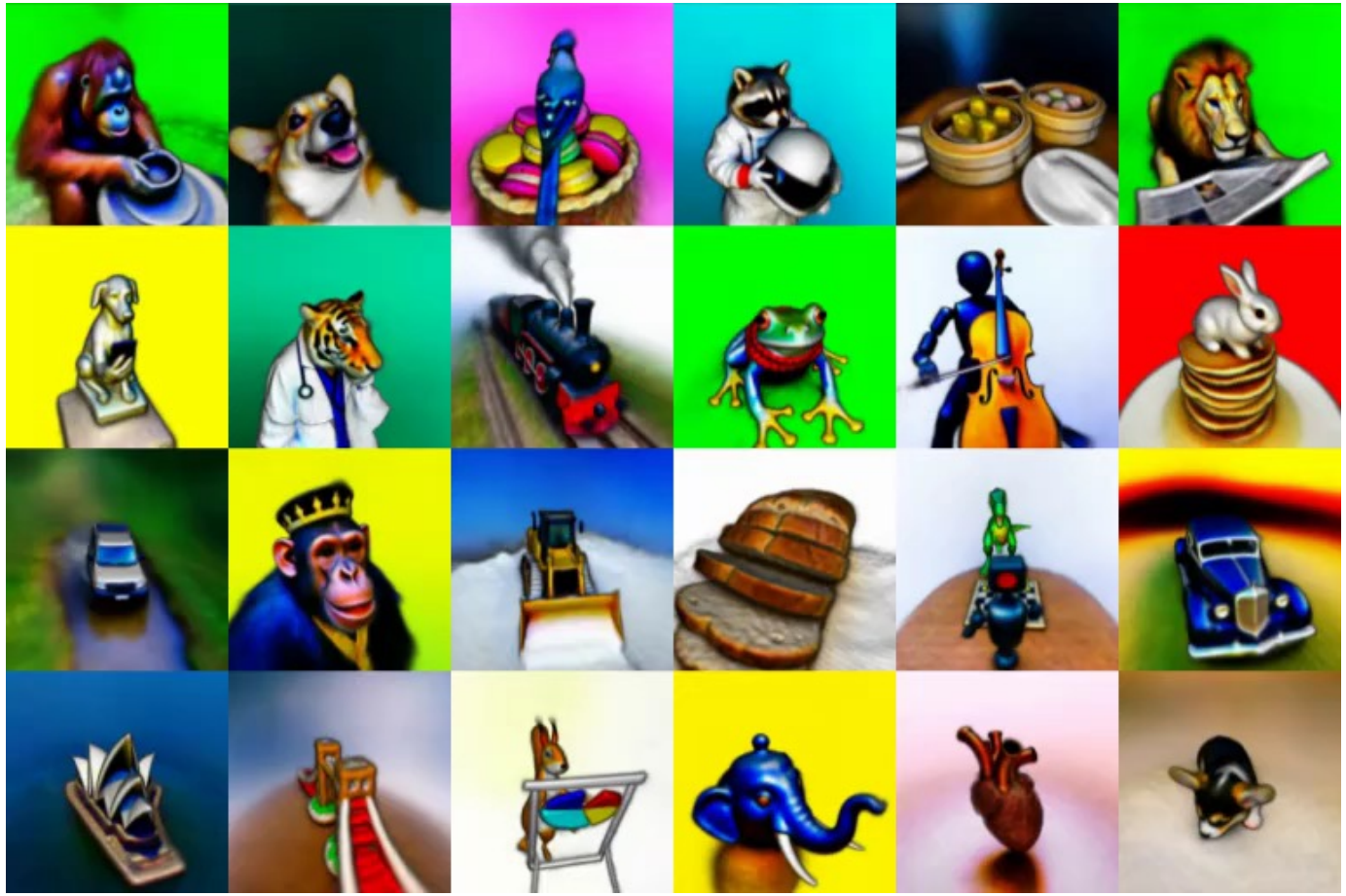
优化三维表示使各个视角渲染图片 “符合”先验分布，无需3D数据

Score distillation sampling（SDS） 目标：渲染图片加入噪声，基础二维扩散模型能准确预测噪声

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\theta) := \mathbb{E}_{t,\epsilon,c}\left[\omega(t)(\epsilon_{\text{pretrain}}(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon})\frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta}\right]$$

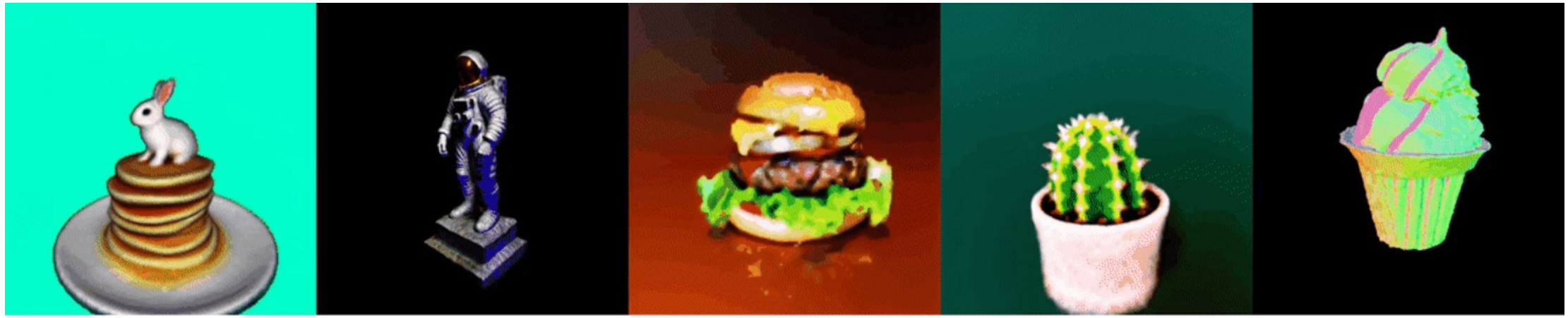
Poole et al. DREAMFUSION: TEXT-TO-3D USING 2D DIFFUSION. ICLR 2023.

DreamFusion：零样本文到三维数据合成结果



DreamFusion 系列工作

- 系列工作尝试提升三维结构表示和基础模型部分，算法均为 **SDS**
- 蒸馏算法是瓶颈：过饱和、过曝、多样性低等现象严重！



Score distillation sampling

- 寻找概率分布的极大值点，而非采样！

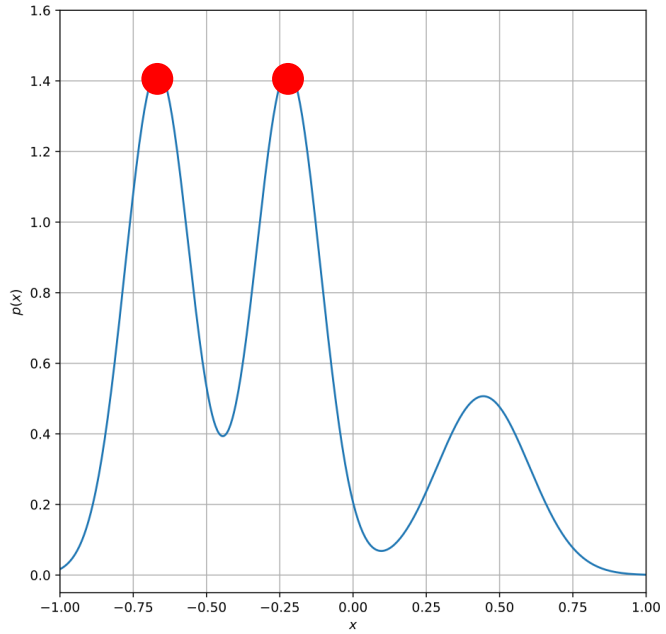
$$\min_{\theta\in\Theta}\mathcal{L}_{\text{SDS}}(\theta) \coloneqq \mathbb{E}_{t,c}\left[(\sigma_t/\alpha_t)\omega(t)D_{\text{KL}}(q_t^\theta(\boldsymbol{x}_t|c)\parallel p_t(\boldsymbol{x}_t|y^c))\right]$$

$$\nabla_\theta\mathcal{L}_{\text{SDS}}(\theta) \triangleq \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t)(\boldsymbol{\epsilon}_{\text{pretrain}}(\boldsymbol{x}_t,t,y)-\boldsymbol{\epsilon})\frac{\partial \boldsymbol{g}(\theta,c)}{\partial\theta}\right]$$

score function　　zero-mean

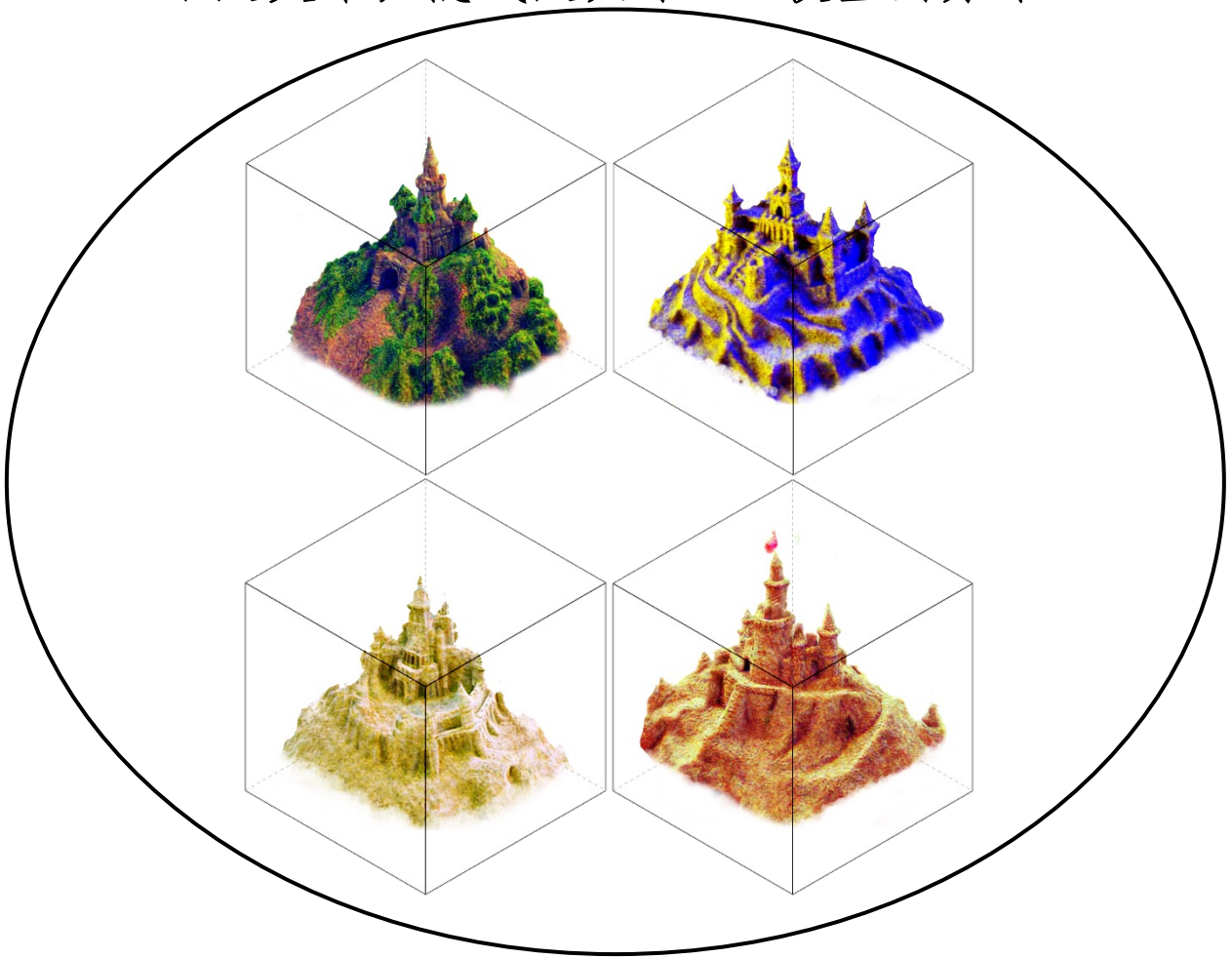
$$\approx \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[-\omega(t)\sigma_t\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t,y)\frac{\partial \boldsymbol{g}(\theta,c)}{\partial\theta}\right]$$

mode-seeking
in 2D space

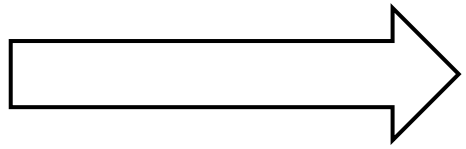


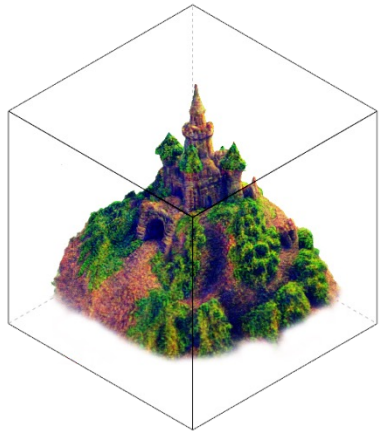
ProlificDreamer：刻画文本到三维内容生成的不确定性

给定文本，隐式定义了 **3D** 模型的分布



从中采样（而非取极值）

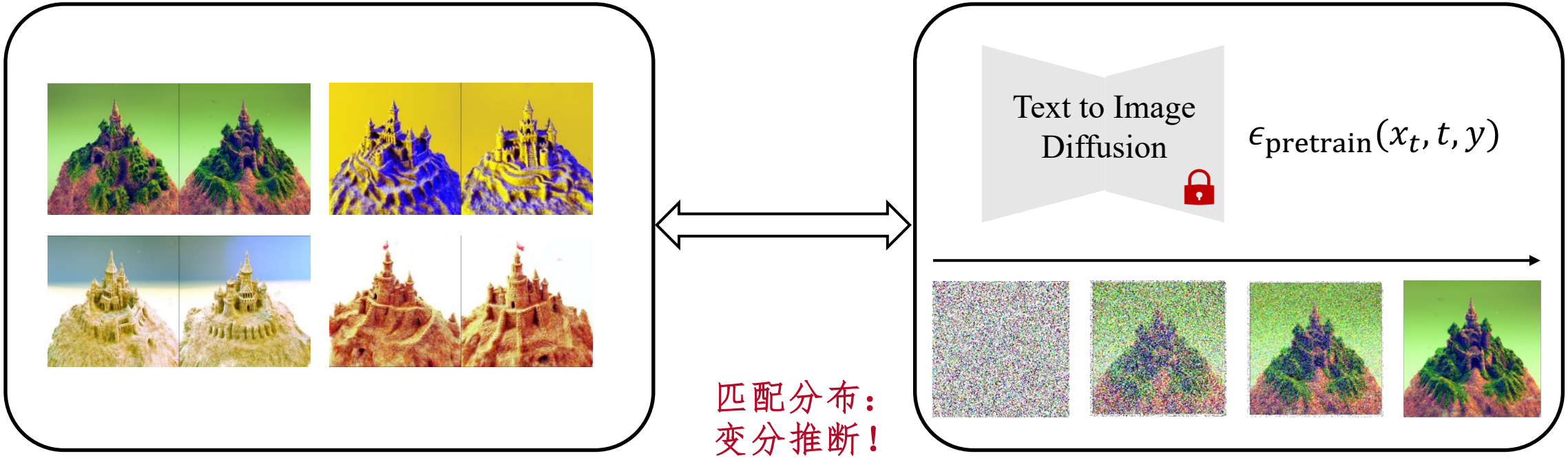




目的：提高保真度和多样性

Wang et al. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. NeurIPS 2023.

文到3D生成：变分推断视角



三维内容分布诱导的渲染图像分布

基础扩散模型定义的图像分布

$$\min KL(\text{渲染图像分布} \| \text{二维模型分布})$$

Variational score distillation（VSD） 算法

- 基于粒子的变分推断保证粒子优化后 **i.i.d.** 二维模型分布

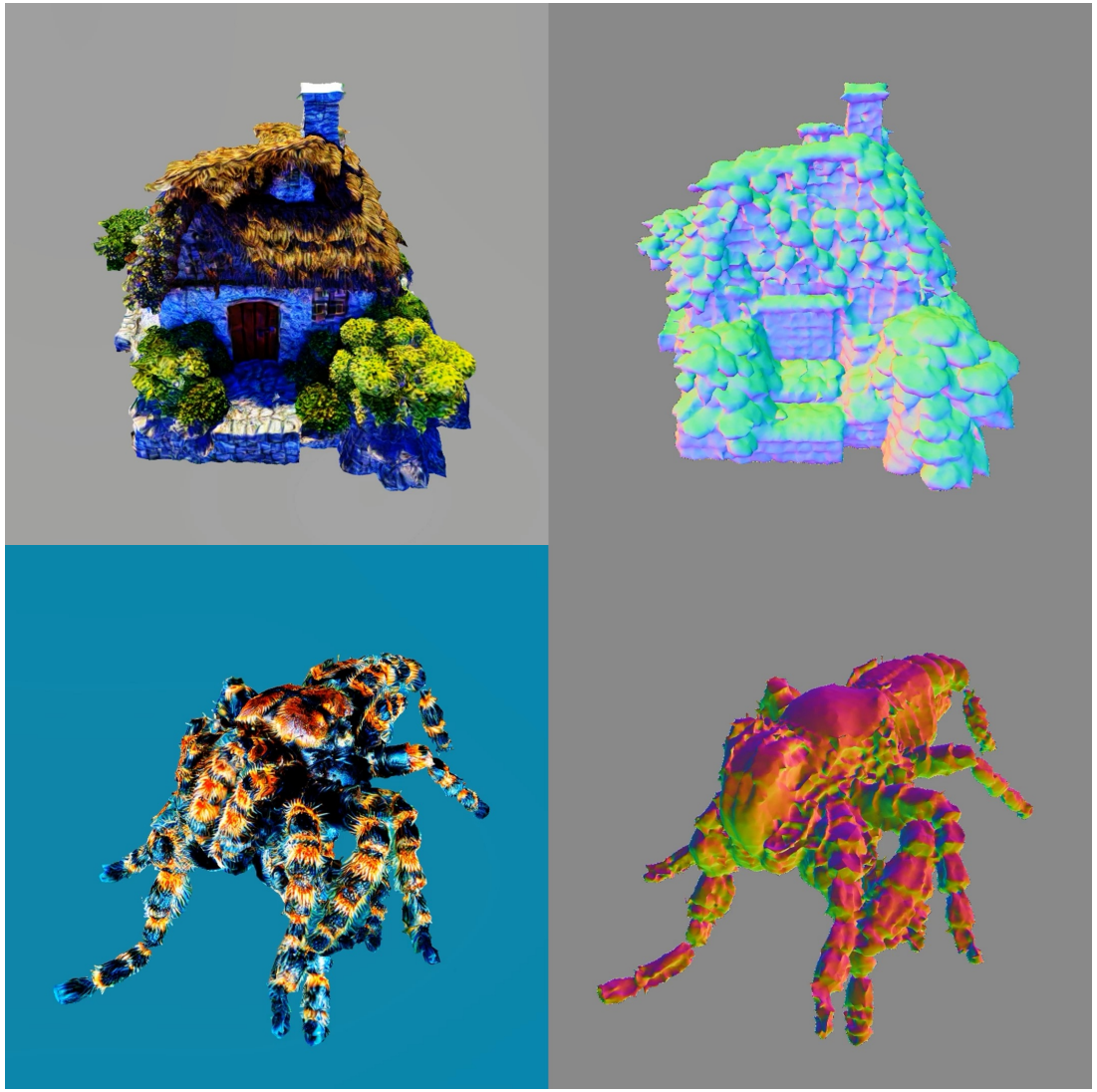
定理：**Wasserstein gradient flow of VSD.**

$$\frac{\mathrm{d}\theta_\tau}{\mathrm{d}\tau} = -\mathbb{E}_{t,\boldsymbol{\epsilon},c}\Big[\omega(t)\Big(\underbrace{-\sigma_t\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|y^c)}_{\text{score of noisy real images}} - \underbrace{(-\sigma_t\nabla_{\boldsymbol{x}_t}\log q_t^{\mu_\tau}(\boldsymbol{x}_t|c,y))}_{\text{score of noisy rendered images}}\Big)\frac{\partial \boldsymbol{g}(\theta_\tau,c)}{\partial\theta_\tau}\Big]$$

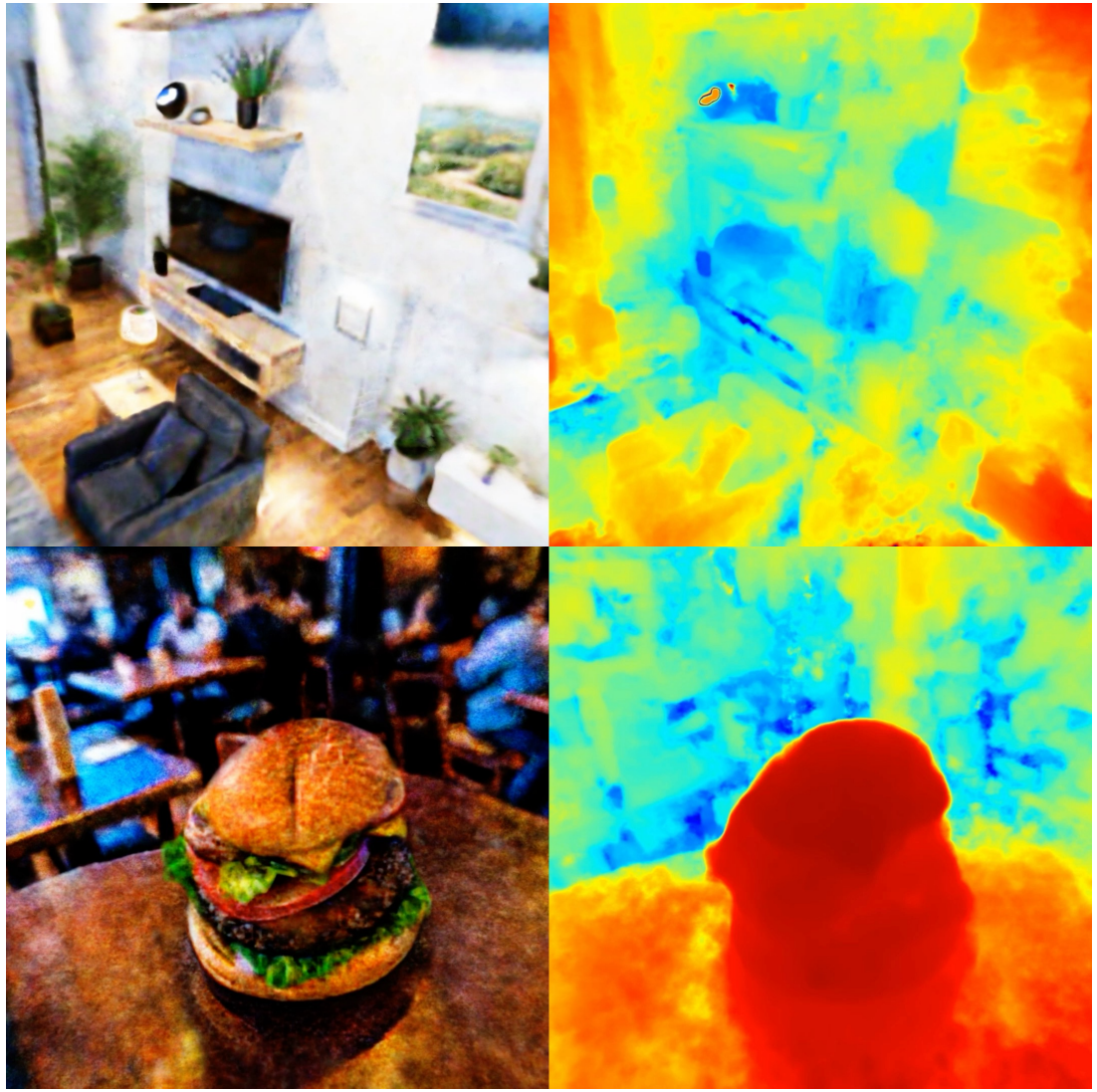
已知二维模型分布　未知渲染图像分布，采用 **LoRA** 估计

VSD 中对应项为已知高斯噪声

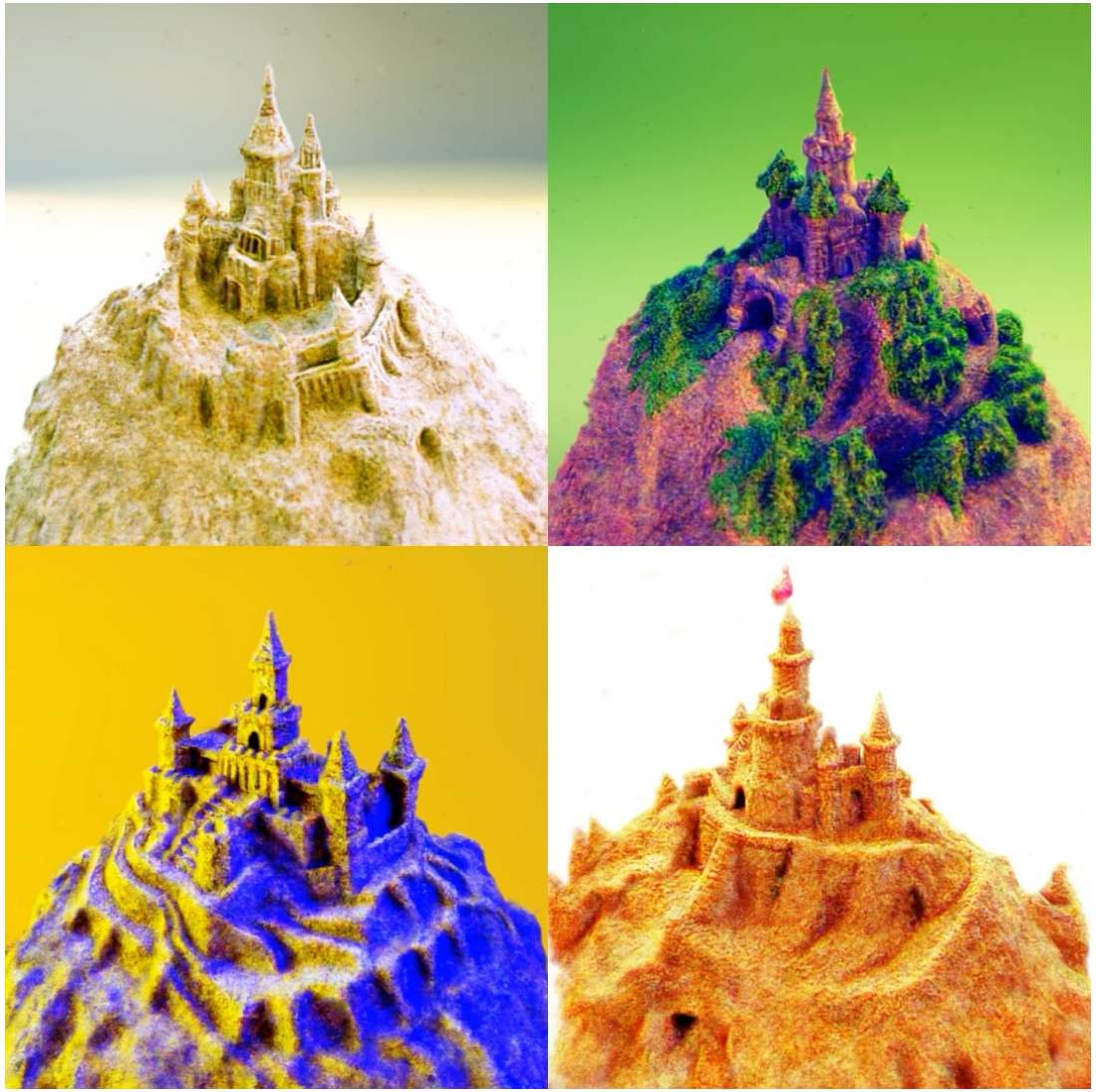
Mesh：高保真度、几何细节丰富



NeRF：高渲染精度、场景、半透明效果



同一个文本的多样性



模态上的新挑战II：可控视频编辑

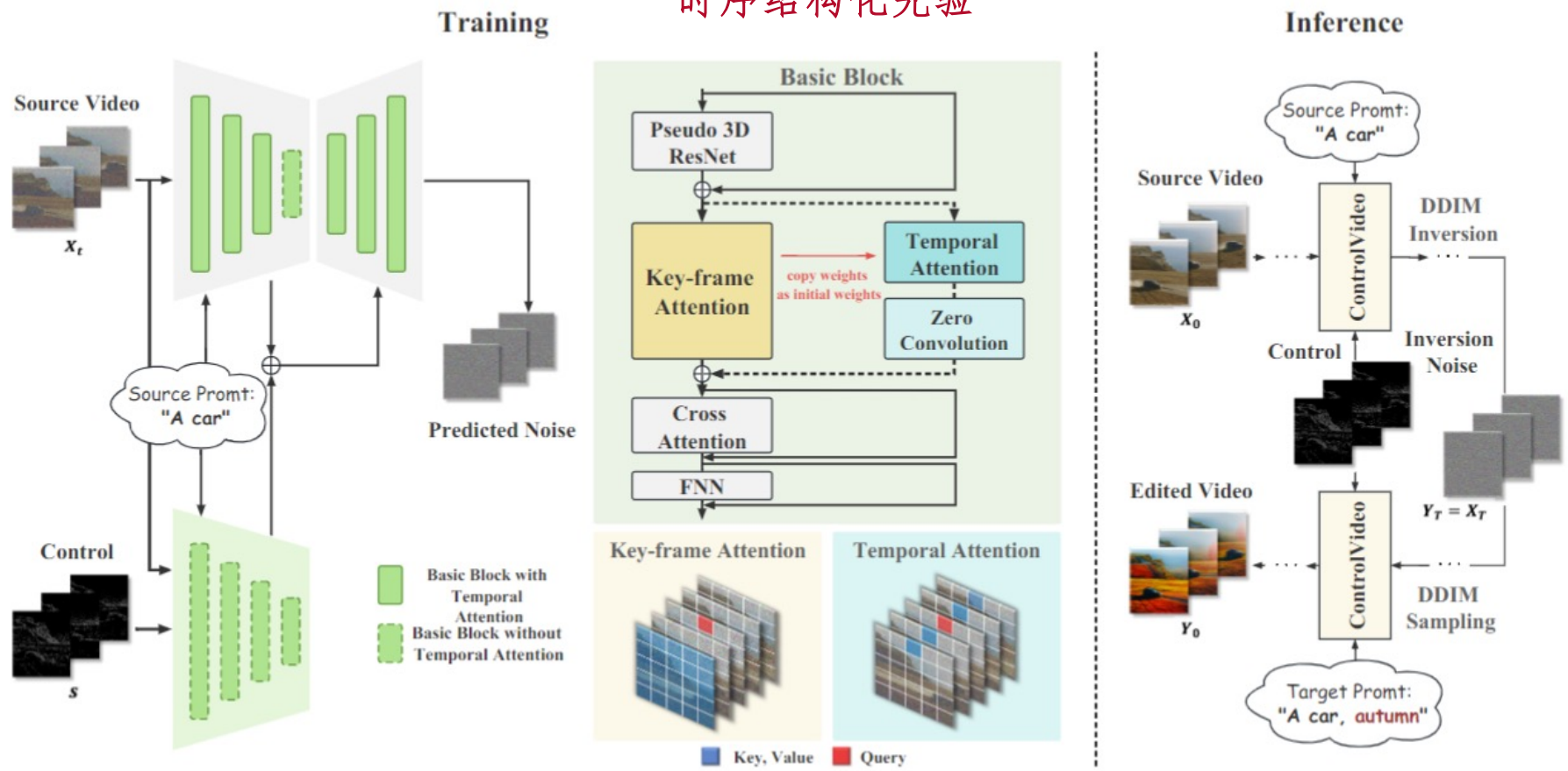
核心：时间一致性

swan→ Swarovski crystal swan

ControlVideo：单样本视频-文本数据做细粒度可控视频编辑

基础扩散模型定义的图像先验

时序结构化先验

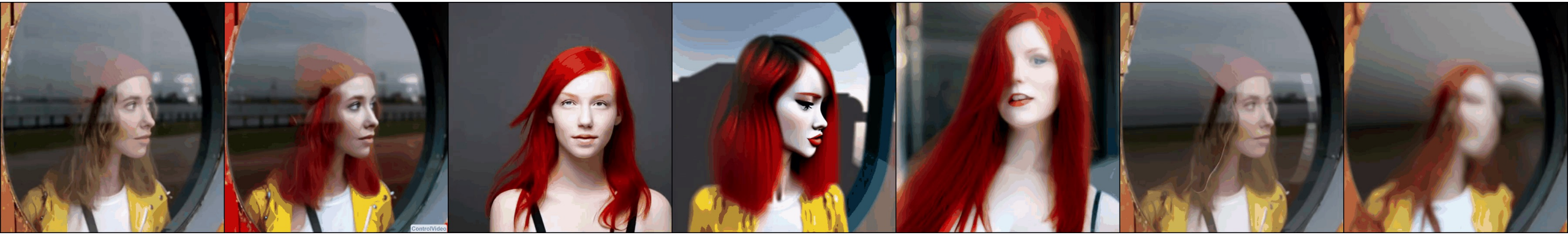


逐帧加入控制条件，加入关键帧和时序注意力机制，精心初始化，单视频微调

Zhao et al. ControlVideo: Adding Conditional Control for One Shot Text-to-Video Editing. 2023.

ControlVideo：结果

Source | **ControlVideo** | SD | Tune-A-Video | Vid2vid-zero | Video-P2P | FateZero



+ with red hair

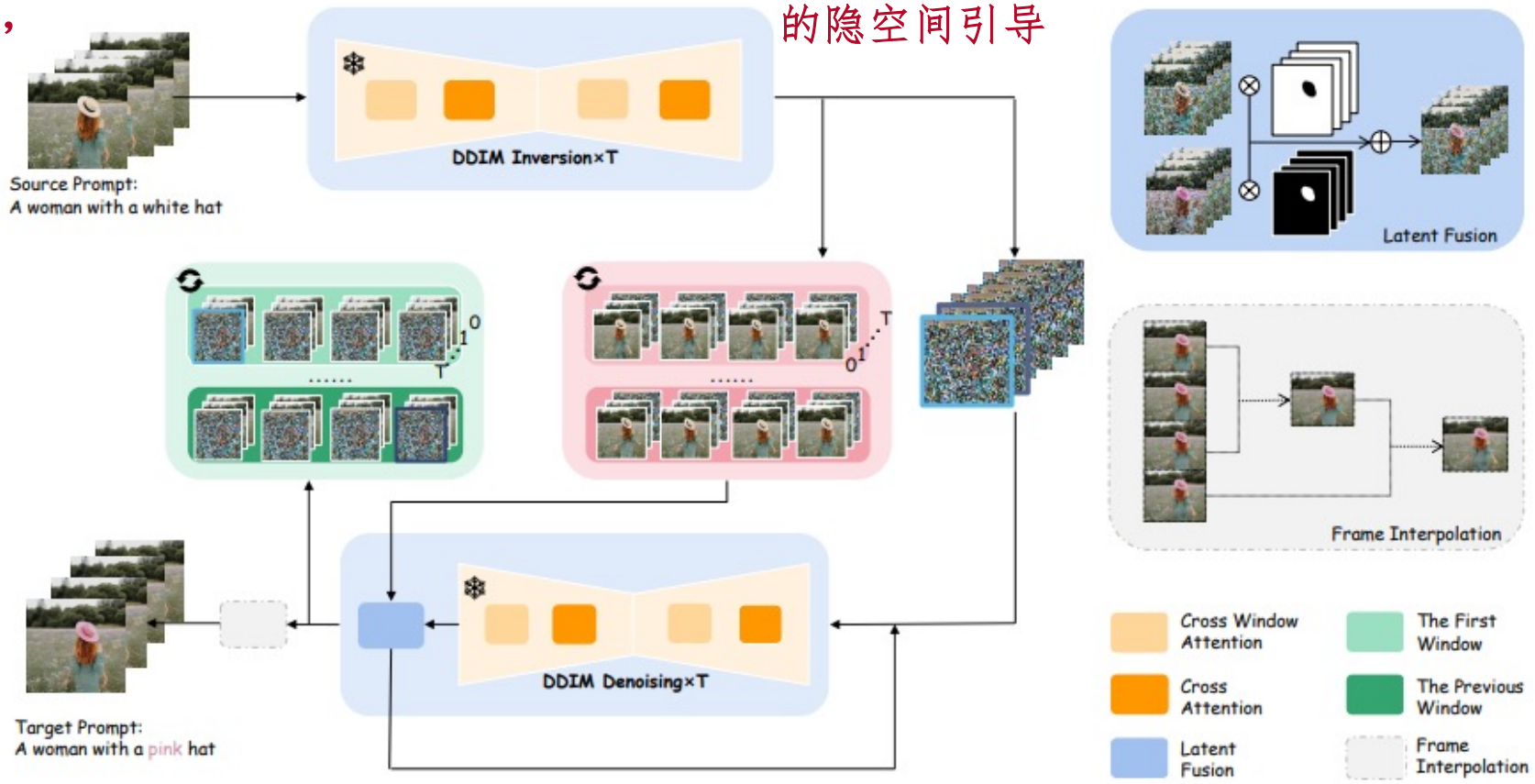
+ Krenz Cushart style

权衡可编辑性、逼真程度、时间一致性、原始内容

LOVECon：基于ControlNet、无需训练、更长视频编辑

分成连续窗口，
逐窗口处理

DDIM Inv. 提供可靠
的隐空间引导



动态融合模块

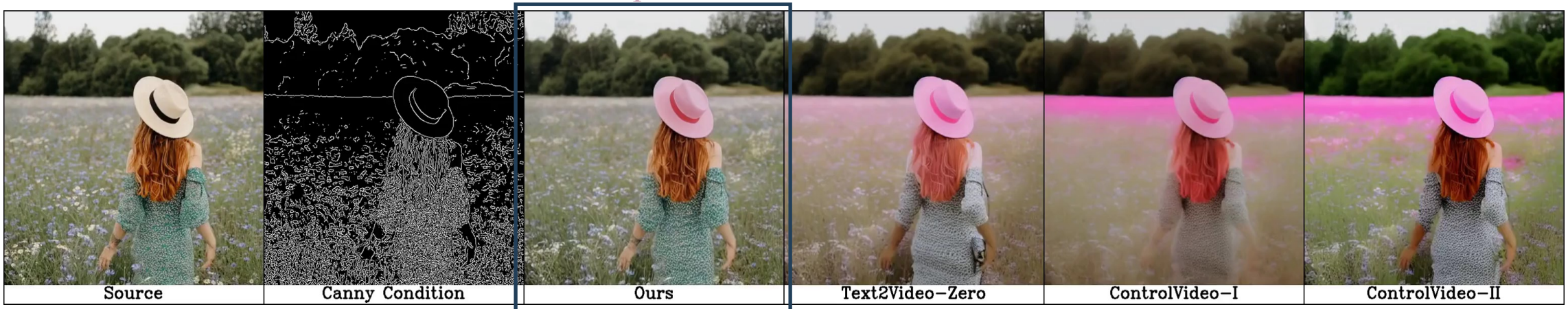
视频插帧模块

将自注意力改造为跨窗口注意力保证连贯生成，将原视频的**DDIM Inv.**状态与生成状态进行动态融合保证无关信息一致，引入视频插帧模型进行后处理提高时序平滑性

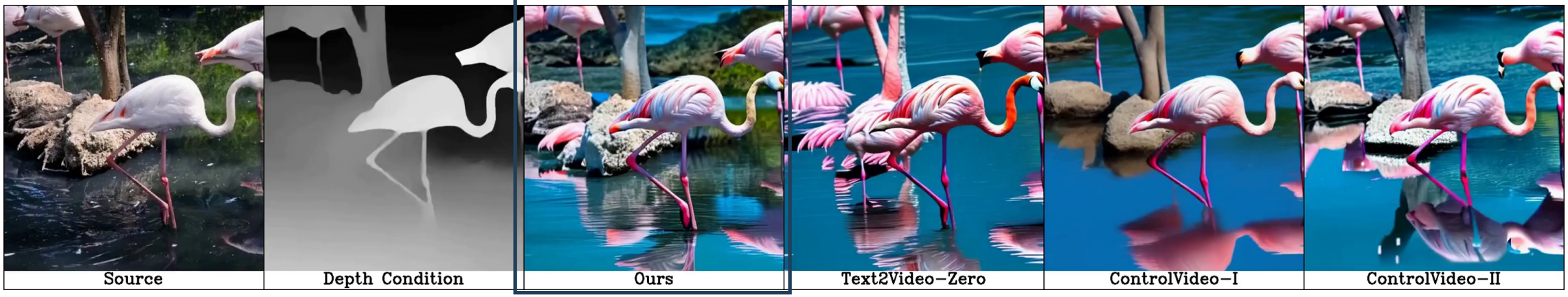
Liao & Deng. LOVECon: Text-driven Training-Free Long Video Editing with ControlNet. 2023.

LOVECon： 结果

A woman with a white hat -> A woman with a pink hat

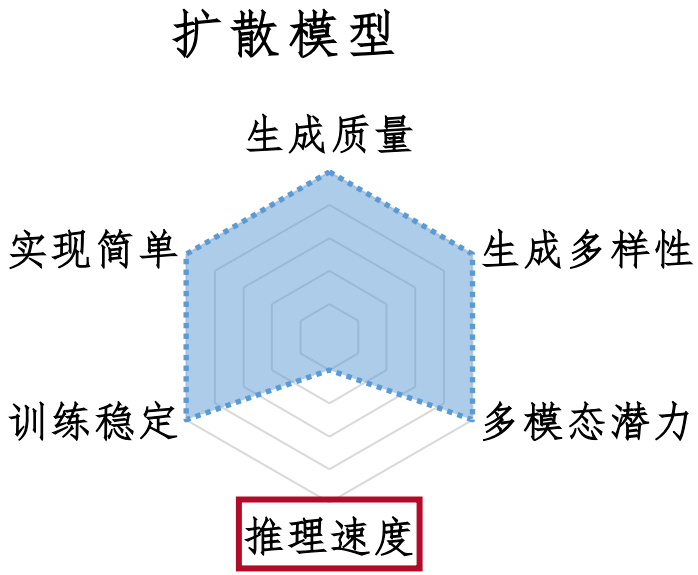


Flamingos -> Flamingos in the blue water

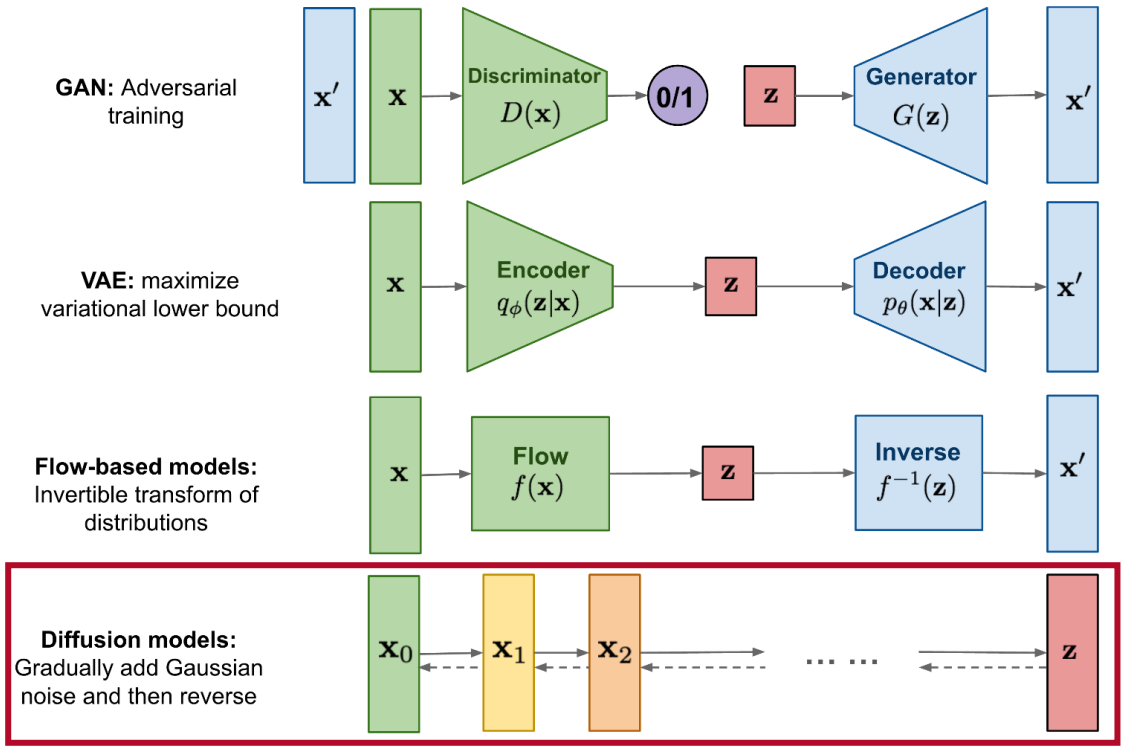


生成的更长、更连贯，与原始内容尽可能一致

采样效率是扩散模型的瓶颈问题

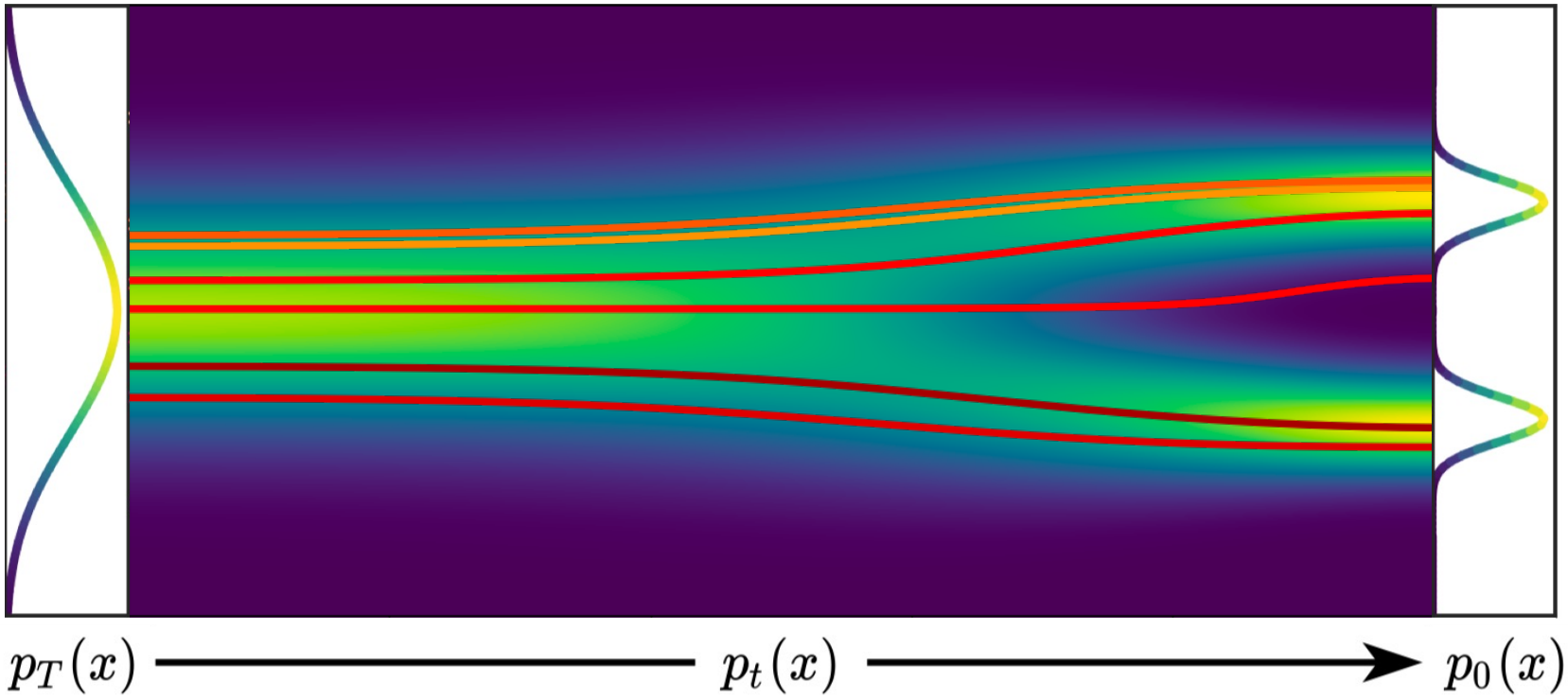


采样效率低：每个去噪步都需要一次模型前传



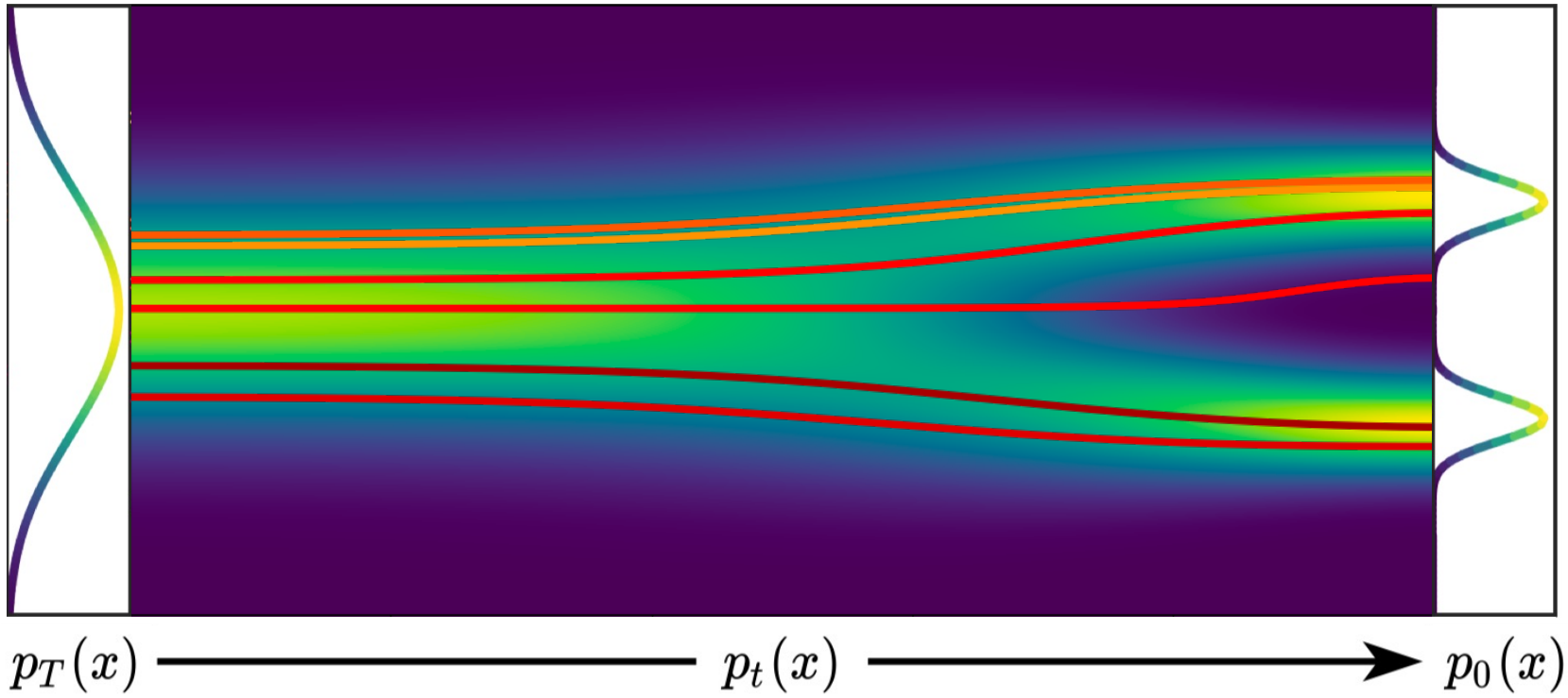
原则上需要迭代1000次，每次需调用神经网络

出发点：常微分方程的视角



时时刻刻边缘分布相同，但是路径上没有噪声！

扩散模型采样等价常微分方程离散化



Sampling method	Steps to converge
Traditional **SDE** Solvers	~200
Traditional **ODE** Solvers	~100

Checkpoint: CIFAR-10, VP

[Song et al., ICLR'21]

DPM-Solver：面向扩散概率模型的常微分方程离散化

- 针对扩散概率模型半线性等特点，设计等价常微分方程的离散化解析形式

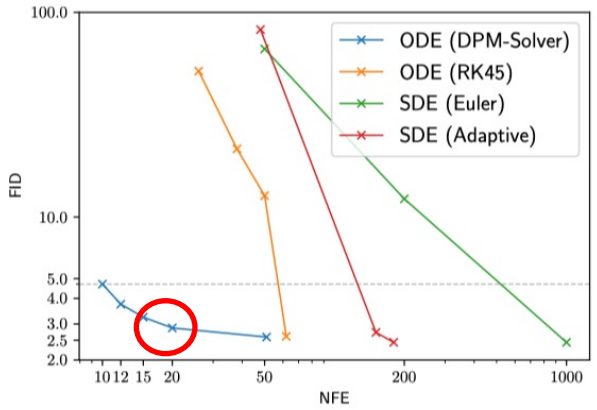
经典龙格库塔法 $$\boldsymbol{x}_t = \boldsymbol{x}_s + \int_s^t \left(f(\tau)\boldsymbol{x}_\tau + \frac{g^2(\tau)}{2\sigma_\tau}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_\tau, \tau) \right) \mathrm{d}\tau$$ 整体黑盒泰勒展开并做差分近似

所提 **DPM-Solver** $$\boldsymbol{x}_{t_{i-1}\to t_i} = \frac{\alpha_{t_i}}{\alpha_{t_{i-1}}}\tilde{\boldsymbol{x}}_{t_{i-1}} - \alpha_{t_i}\sum_{n=0}^{k-1}\hat{\boldsymbol{\epsilon}}_\theta^{(n)}(\hat{\boldsymbol{x}}_{\lambda_{t_{i-1}}}, \lambda_{t_{i-1}})\int_{\lambda_{t_{i-1}}}^{\lambda_{t_i}} e^{-\lambda}\frac{(\lambda-\lambda_{t_{i-1}})^n}{n!}\mathrm{d}\lambda + \mathcal{O}(h_i^{k+1})$$

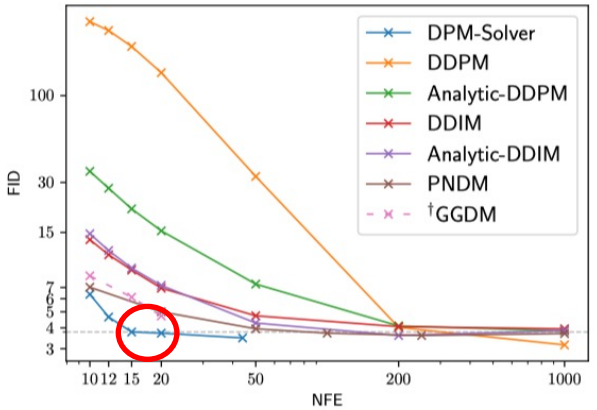
解析形式　　神经网络部分差分近似　　解析形式　　高阶小量

Lu C, et al. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. NeurIPS 2022.

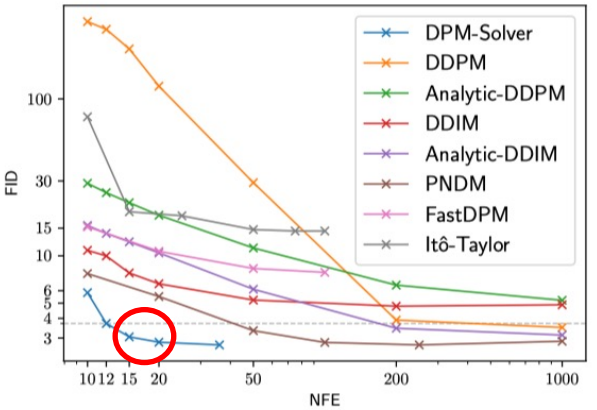
DPM-Solver：结果



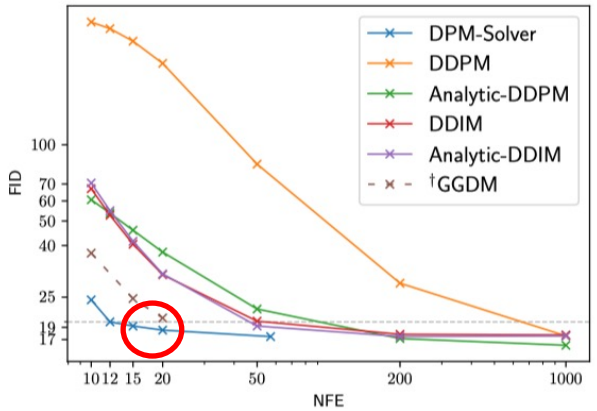
(a) CIFAR-10 (continuous)



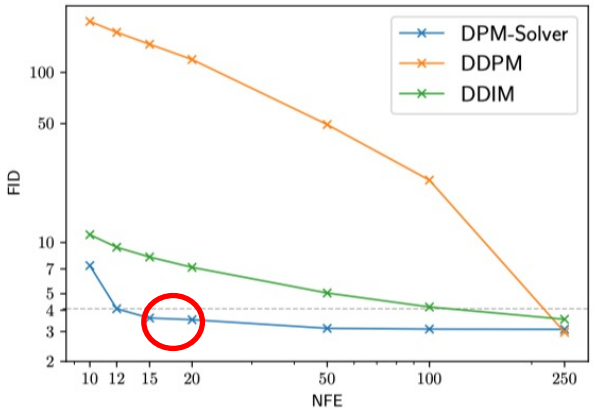
(b) CIFAR-10 (discrete)



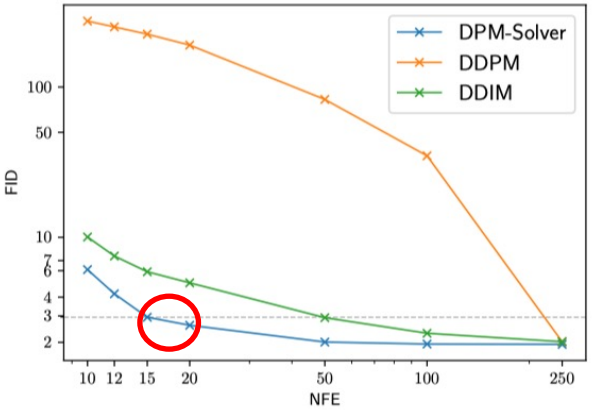
(c) CelebA 64x64 (discrete)



(d) ImageNet 64x64 (discrete)



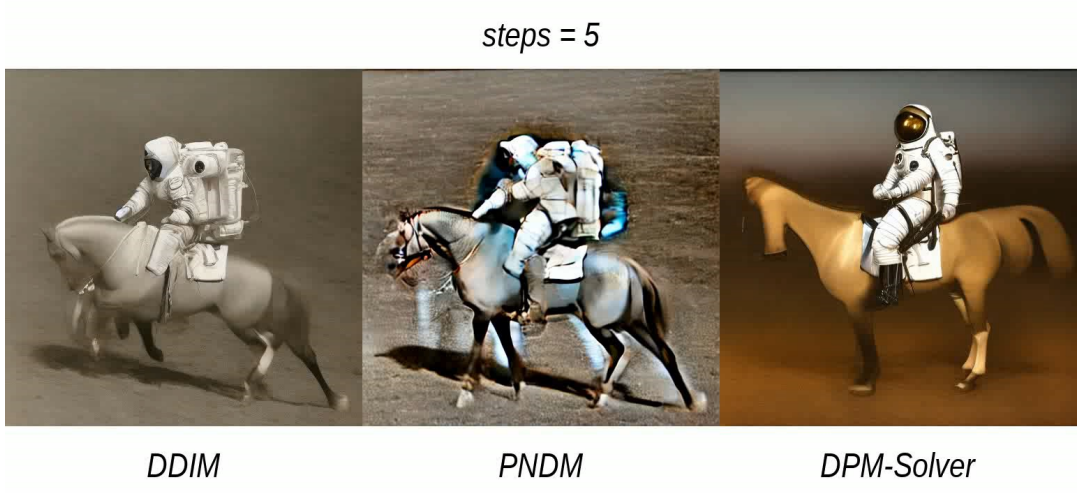
(e) ImageNet 128x128 (discrete)



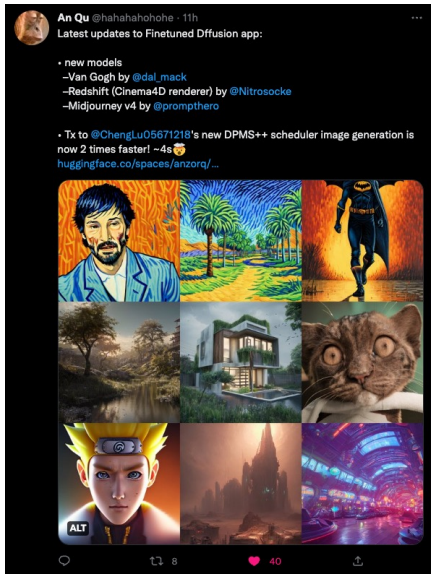
(f) LSUN bedroom 256x256 (discrete)

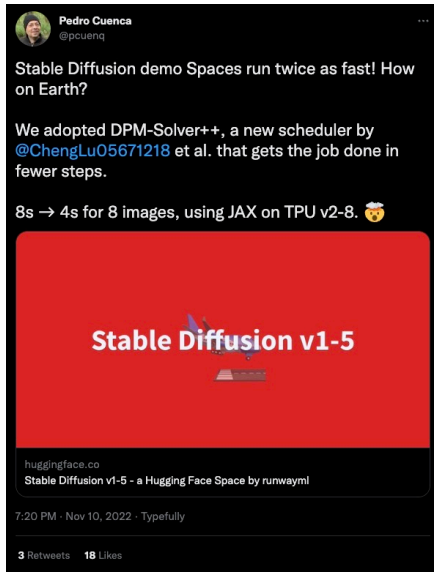
DPM-Solver：结果

- 是当前最快的无需额外学习的扩散概率模型采样算法，**15 步生成高清图像**
- 被多个主流开源社区（**Github** 累计星标 **6** 万余次）支持/设为默认算法



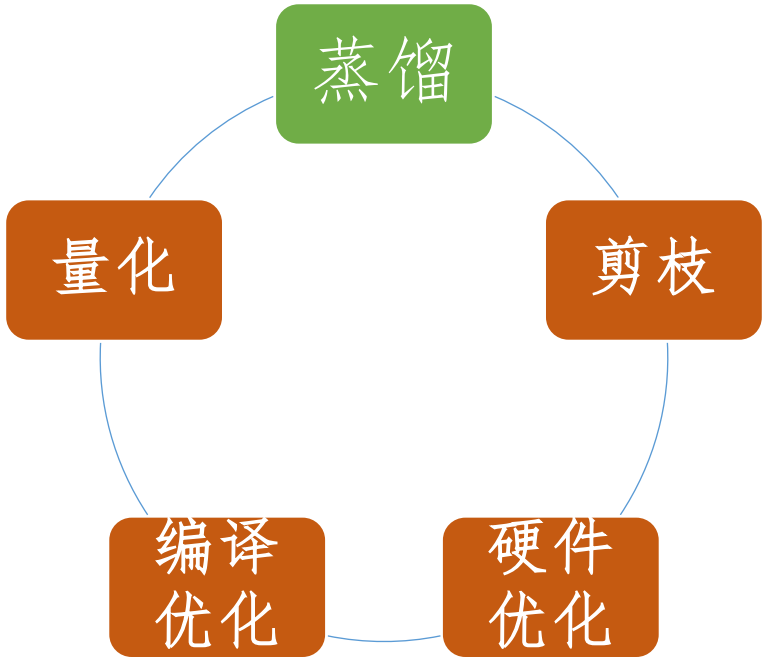
根据文本输入 **15** 步生成 **512x512** 高清图像



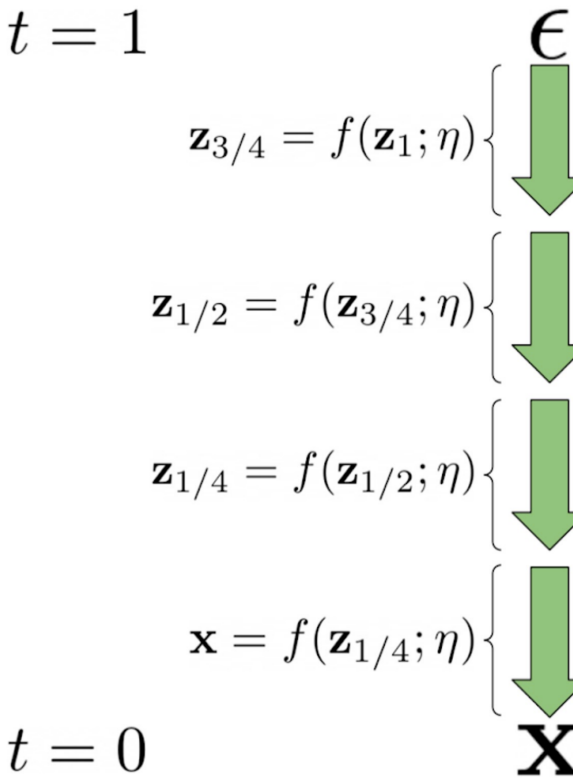


著名开源模型 **Stable Diffusion** 等官方宣传

从深度学习模型加速的视角出发



Progressive Distillat



Distillat

Algorithm 2 Progressive distillation

Require: Trained teacher model $\hat{\mathbf{x}}_\eta(\mathbf{z}_t)$
Require: Data set $\mathcal{D}$
Require: Loss weight function $w()$
Require: Student sampling steps N
  for K iterations **do**
    $\theta \leftarrow \eta$ ▷ Init student from teacher
    while not converged **do**
      $\mathbf{x} \sim \mathcal{D}$
      $t = i/N, \;\; i \sim Cat[1, 2, \ldots, N]$
      $\epsilon \sim N(0, I)$
      $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$
      `# 2 steps of DDIM with teacher`
      $t' = t - 0.5/N, \;\; t'' = t - 1/N$
      $\mathbf{z}_{t'} = \alpha_{t'} \hat{\mathbf{x}}_\eta(\mathbf{z}_t) + \frac{\sigma_{t'}}{\sigma_t}(\mathbf{z}_t - \alpha_t \hat{\mathbf{x}}_\eta(\mathbf{z}_t))$
      $\mathbf{z}_{t''} = \alpha_{t''} \hat{\mathbf{x}}_\eta(\mathbf{z}_{t'}) + \frac{\sigma_{t''}}{\sigma_{t'}}(\mathbf{z}_{t'} - \alpha_{t'} \hat{\mathbf{x}}_\eta(\mathbf{z}_{t'}))$
      $\tilde{\mathbf{x}} = \frac{\mathbf{z}_{t''} - (\sigma_{t''}/\sigma_t)\mathbf{z}_t}{\alpha_{t''} - (\sigma_{t''}/\sigma_t)\alpha_t}$ ▷ Teacher $\hat{\mathbf{x}}$ target
      $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$
      $L_\theta = w(\lambda_t)\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t)\|_2^2$
      $\theta \leftarrow \theta - \gamma \nabla_\theta L_\theta$
    end while
    $\eta \leftarrow \theta$ ▷ Student becomes next teacher
    $N \leftarrow N/2$ ▷ Halve number of sampling steps
  end for

提前确定noise schedule

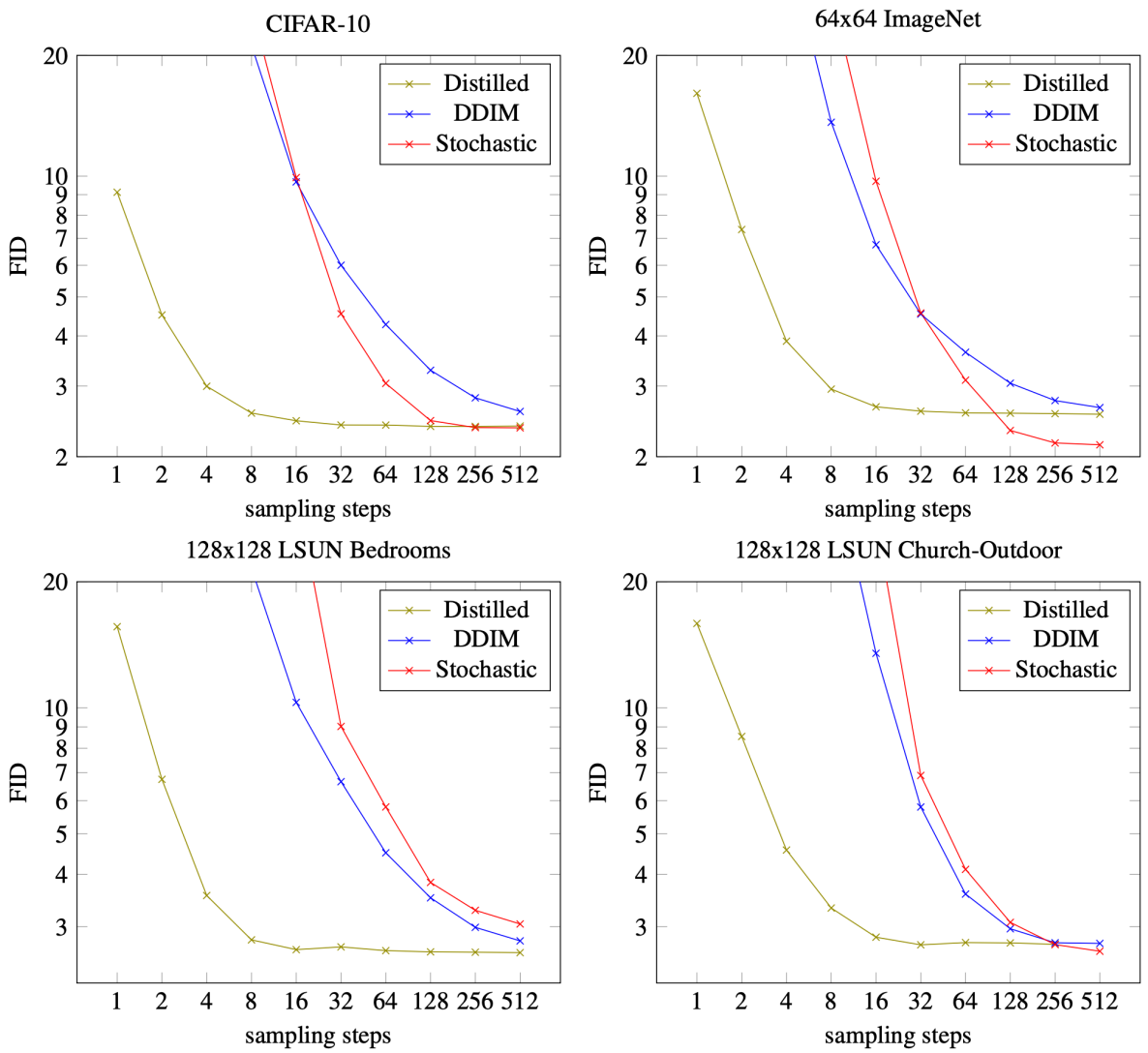
基于预训练模型，以**确定性ODE sampler作为教师**

2步蒸馏为1步：学生只需要教师的一半采样步

循环往复

Salimans & Ho. Progressive Distillation for Fast S

Progressive Distillation：结果



在不显著影响生成质量的情况下，将需要的采样步**降低到4**：

- CIFAR-10 上FID：3.0

Salimans & Ho. Progressive Distillation for Fast Sampling of Diffusion Models. ICLR 2022

改进：On Distillation of Guided Diffusion Models (Meng et al., CVPR 2023)

- 带来了：
  - CF-Guidance
  - Stochastic sampling
  - Text-to-image
  - Image-to-image
  - Inpainting
  - Latent Diffusion

Text-guided generation (1 step)

Text-guided generation (2 steps)

Class-conditional generation (1 step)

Input Mask Result 1 Result 2

Image inpainting (2 steps)

Text-guided generation (4 steps)

Input Output (different styles)

Image to image translation (3 steps)

1-4步即可生成高质量大图

蒸馏的终极目标：一步生成

Consistency Distillation

- 概率流ODE定义了从噪声到数据的一一映射

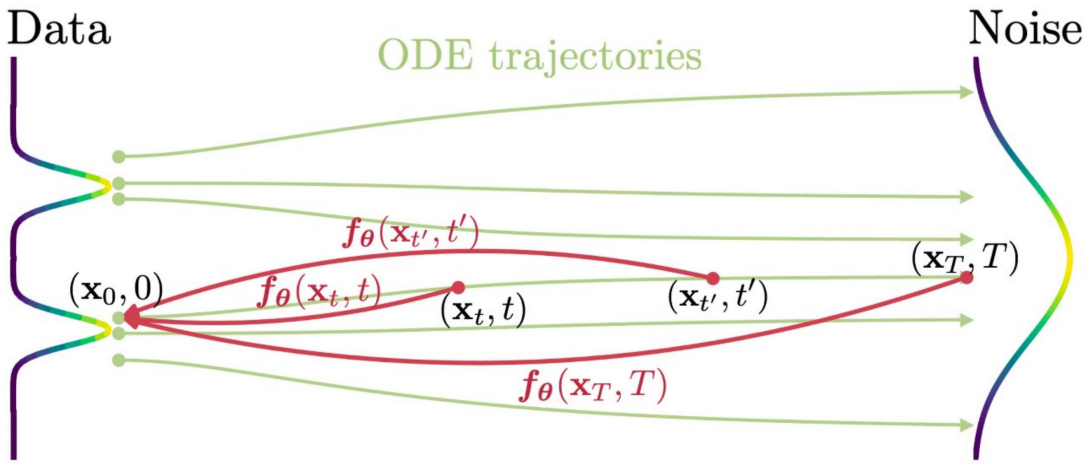
  =>直接建模此映射

$$\forall t \in [0, T] : \boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \mathbf{x}_0$$

- **模型参数化：需保证边界条件（t=0）**

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}, t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t)F_{\boldsymbol{\theta}}(\mathbf{x}, t)$$

$$c_{\text{skip}}(0) = 1 \qquad c_{\text{out}}(0) = 0$$



- **训练：**

One ODE solver step using pretrained diffusion model

$$\mathbb{E}\left[\lambda(t_n)\left\|\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_{t_{n+1}}, t_{n+1}) - \boldsymbol{f}_{\boldsymbol{\theta}^-}(\hat{\mathbf{x}}_{t_n}, t_n)\right\|\right]$$

Weighting function　Student model　Teacher model　$\boldsymbol{\theta}^- = \text{EMA}(\boldsymbol{\theta})$

Algorithm 1 Multistep Consistency Sampling

Input: Consistency model $\boldsymbol{f}_{\boldsymbol{\theta}}(\cdot,\cdot)$, sequence of time points $\tau_1 > \tau_2 > \cdots > \tau_{N-1}$, initial noise $\hat{\mathbf{x}}_T$

$\mathbf{x} \leftarrow \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_T, T)$

for $n = 1$ **to** $N - 1$ **do**

  Sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

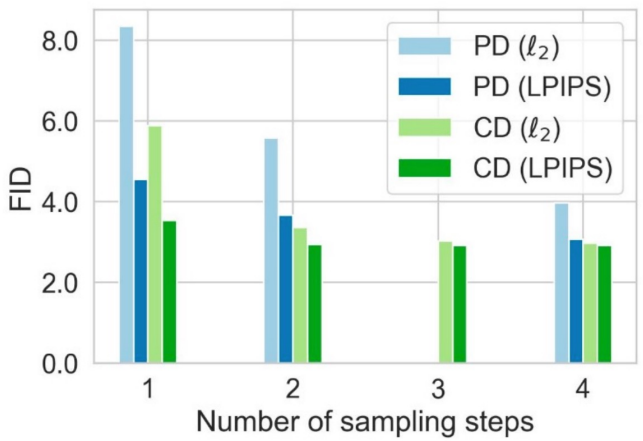
  $\hat{\mathbf{x}}_{\tau_n} \leftarrow \mathbf{x} + \sqrt{\tau_n^2 - \epsilon^2}\mathbf{z}$

  $\mathbf{x} \leftarrow \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_{\tau_n}, \tau_n)$

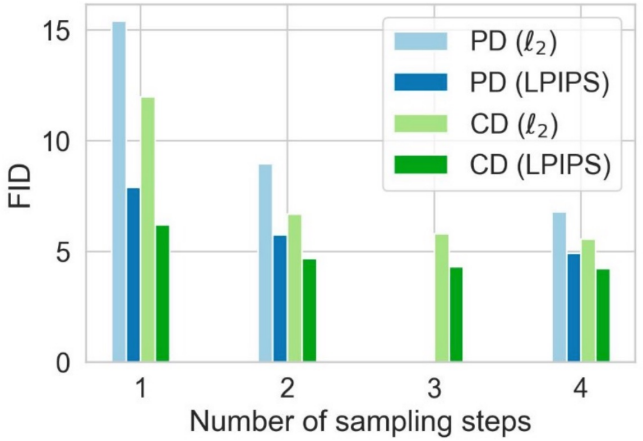
end for

Output: $\mathbf{x}$

Consistency Distillation：结果



(a) CIFAR-10

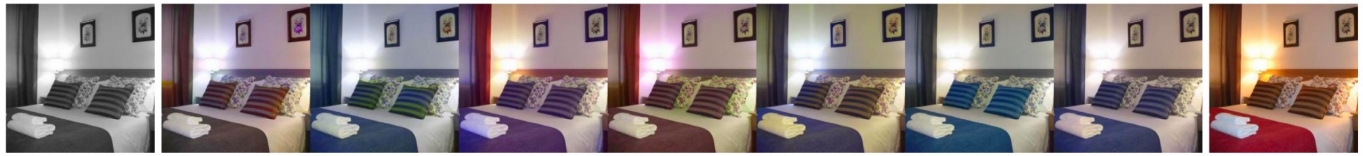


(b) ImageNet 64 × 64

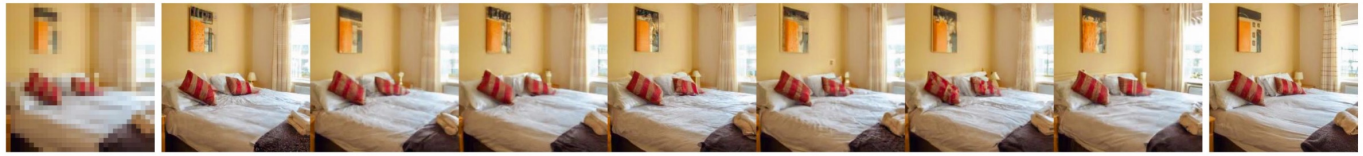
一步生成的**SOTA** FID:

- 3.55 on CIFAR-10
- 6.20 on ImageNet 64

可用于**零样本图像编辑**等应用



(a) *Left*: The gray-scale image. *Middle*: Colorized images. *Right*: The ground-truth image.



(b) *Left*: The downsampled image (32 × 32). *Middle*: Full resolution images (256 × 256). *Right*: The ground-truth image (256 × 256).

(c) *Left*: A stroke input provided by users. *Right*: Stroke-guided image generation.

Consistency Distillation：用于DALL-E 3

In a fantastical setting, a highly detailed furry humanoid skunk with piercing eyes confidently poses in a medium shot, wearing an animal hide jacket. The artist has masterfully rendered the character in digital art, capturing the intricate details of fur and clothing texture.

A illustration from a graphic novel. A bustling city street under the shine of a full moon. The sidewalks bustling with pedestrians enjoying the nightlife. At the corner stall, a young woman with fiery red hair, dressed in a signature velvet cloak, is haggling with the grumpy old vendor. the grumpy vendor, a tall, sophisticated man is wearing a sharp suit, sports a noteworthy moustache is animatedly conversing on his steampunk telephone.

https://cdn.openai.com/papers/dall-e-3.pdf

隐空间的Consistency Distillation

适用于蒸馏SD

参数化： $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{c}, t) = c_{\text{skip}}(t)\boldsymbol{z} + c_{\text{out}}(t)\left(\frac{\boldsymbol{z} - \sigma_t \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}, \boldsymbol{c}, t)}{\alpha_t}\right)$, ($\epsilon$-Prediction)

Algorithm 1 Latent Consistency Distillation (LCD)

Input: dataset $\mathcal{D}$, initial model parameter $\boldsymbol{\theta}$, learning rate η, ODE solver $\Psi(\cdot,\cdot,\cdot,\cdot)$, distance metric $d(\cdot,\cdot)$, EMA rate μ, noise schedule $\alpha(t), \sigma(t)$, guidance scale $[w_{\text{min}}, w_{\text{max}}]$, skipping interval k, and encoder $E(\cdot)$

Encoding training data into latent space: $\mathcal{D}_z = \{(\boldsymbol{z}, \boldsymbol{c}) | \boldsymbol{z} = E(\boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{c}) \in \mathcal{D}\}$

引入跳跃步数，但ODE solver走一步

$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$

repeat

Sample $(\boldsymbol{z}, \boldsymbol{c}) \sim \mathcal{D}_z$, $n \sim \mathcal{U}[1, N-k]$ and $\omega \sim [\omega_{\text{min}}, \omega_{\text{max}}]$

Sample $\boldsymbol{z}_{t_{n+k}} \sim \mathcal{N}(\alpha(t_{n+k})\boldsymbol{z}; \sigma^2(t_{n+k})\mathbf{I})$

$\hat{\boldsymbol{z}}^{\Psi,\omega}_{t_n} \leftarrow \boldsymbol{z}_{t_{n+k}} + (1+\omega)\Psi(\boldsymbol{z}_{t_{n+k}}, t_{n+k}, t_n, \boldsymbol{c}) - \omega\Psi(\boldsymbol{z}_{t_{n+k}}, t_{n+k}, t_n, \varnothing)$ 引入CFG

$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \Psi) \leftarrow d(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}_{t_{n+k}}, \omega, \boldsymbol{c}, t_{n+k}), \boldsymbol{f}_{\boldsymbol{\theta}^-}(\hat{\boldsymbol{z}}^{\Psi,\omega}_{t_n}, \omega, \boldsymbol{c}, t_n))$ 输入更多参数

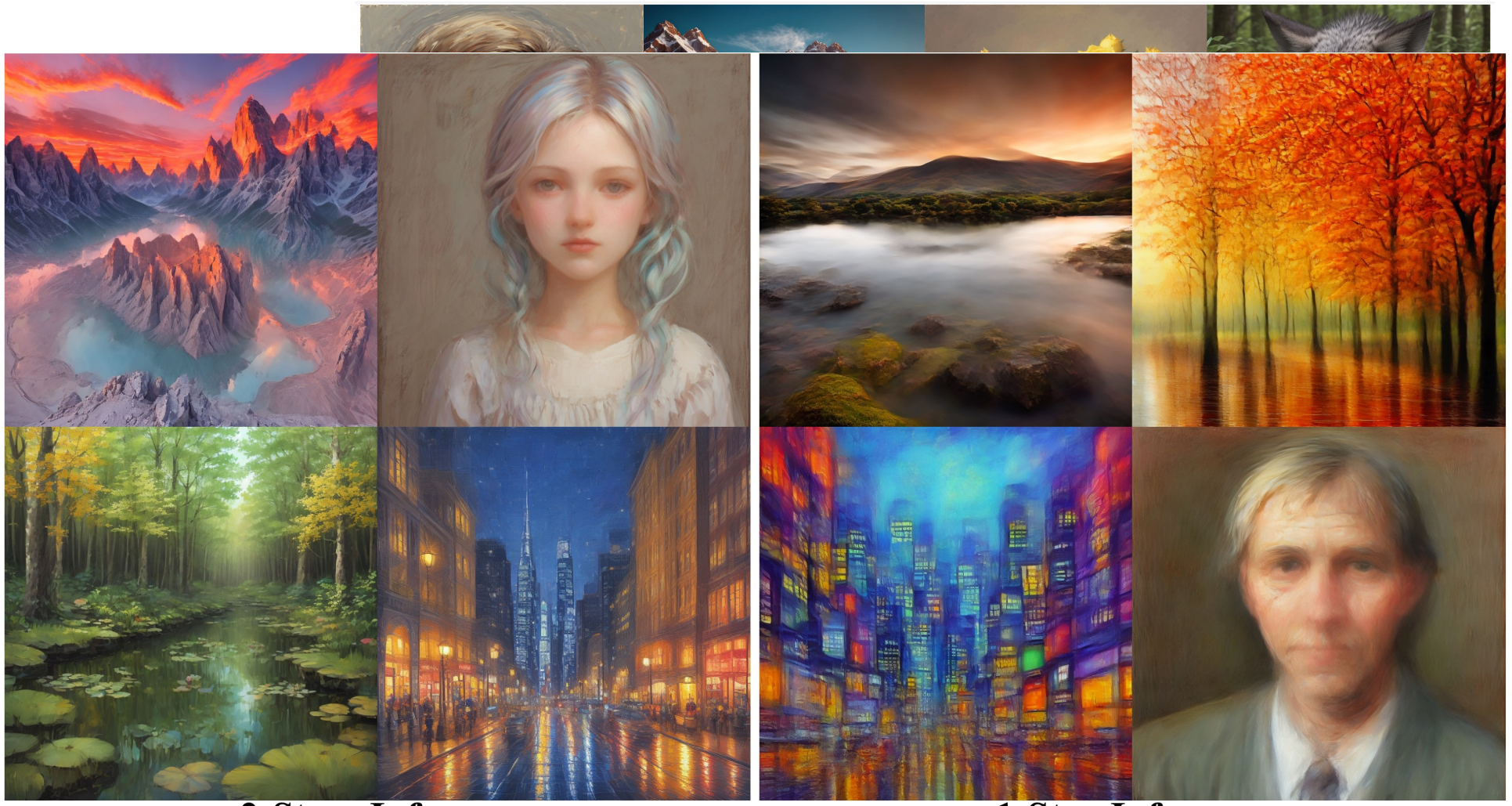
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$

$\boldsymbol{\theta}^- \leftarrow \text{stopgrad}(\mu\boldsymbol{\theta}^- + (1-\mu)\boldsymbol{\theta})$

until convergence

Luo et al. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference.

隐空间的Consistency Distillation



从预训练SD蒸馏：

- 4,000个训练步（约32个A100 GPU hour）
- 生成高质量的768×768分辨率图像

Luo et al. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference.

对抗性扩散蒸馏

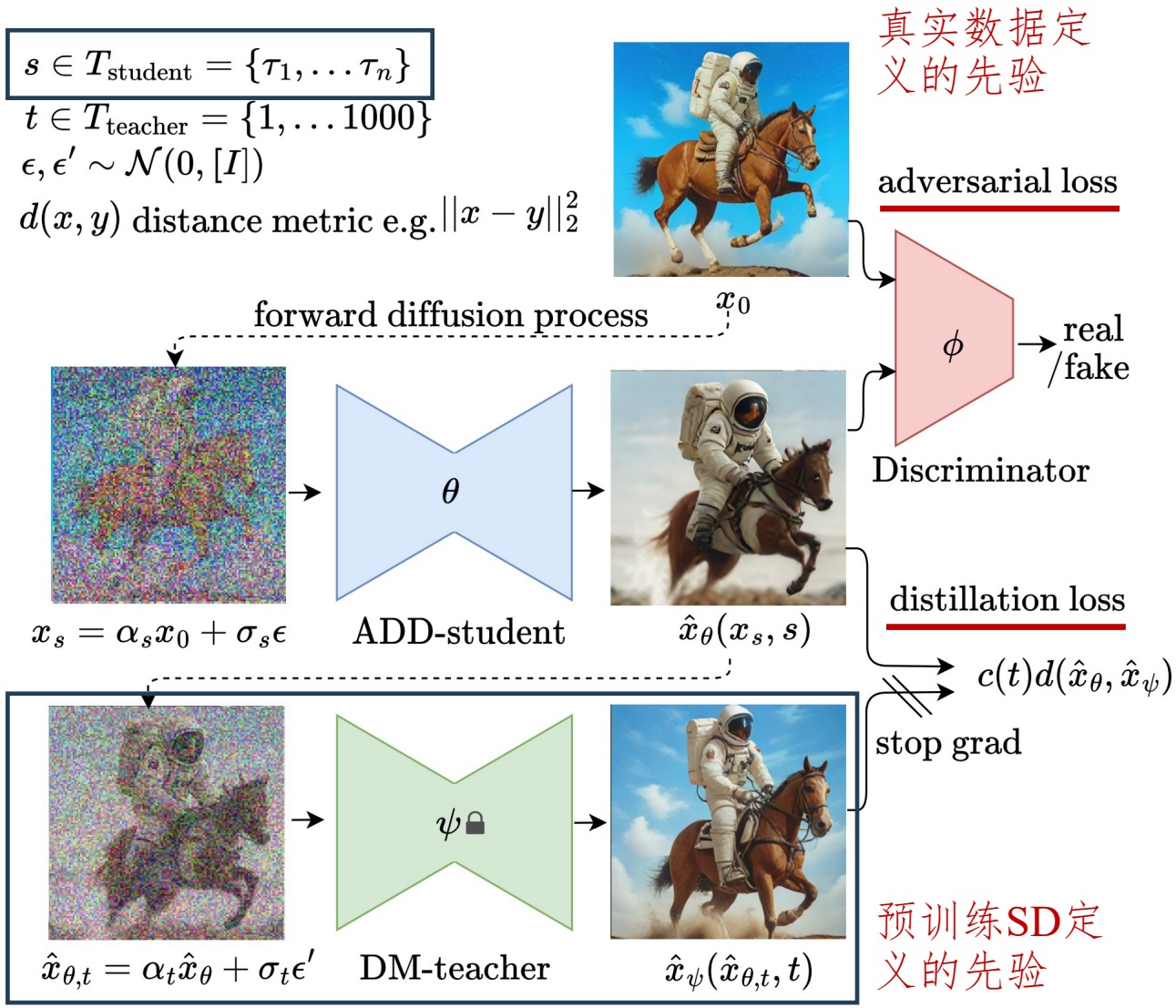
学生只学习4个时间步上的去噪

$s \in T_{\text{student}} = \{\tau_1, \ldots \tau_n\}$

$t \in T_{\text{teacher}} = \{1, \ldots 1000\}$

$\epsilon, \epsilon' \sim \mathcal{N}(0, [I])$

$d(x, y)$ distance metric e.g. $||x - y||_2^2$

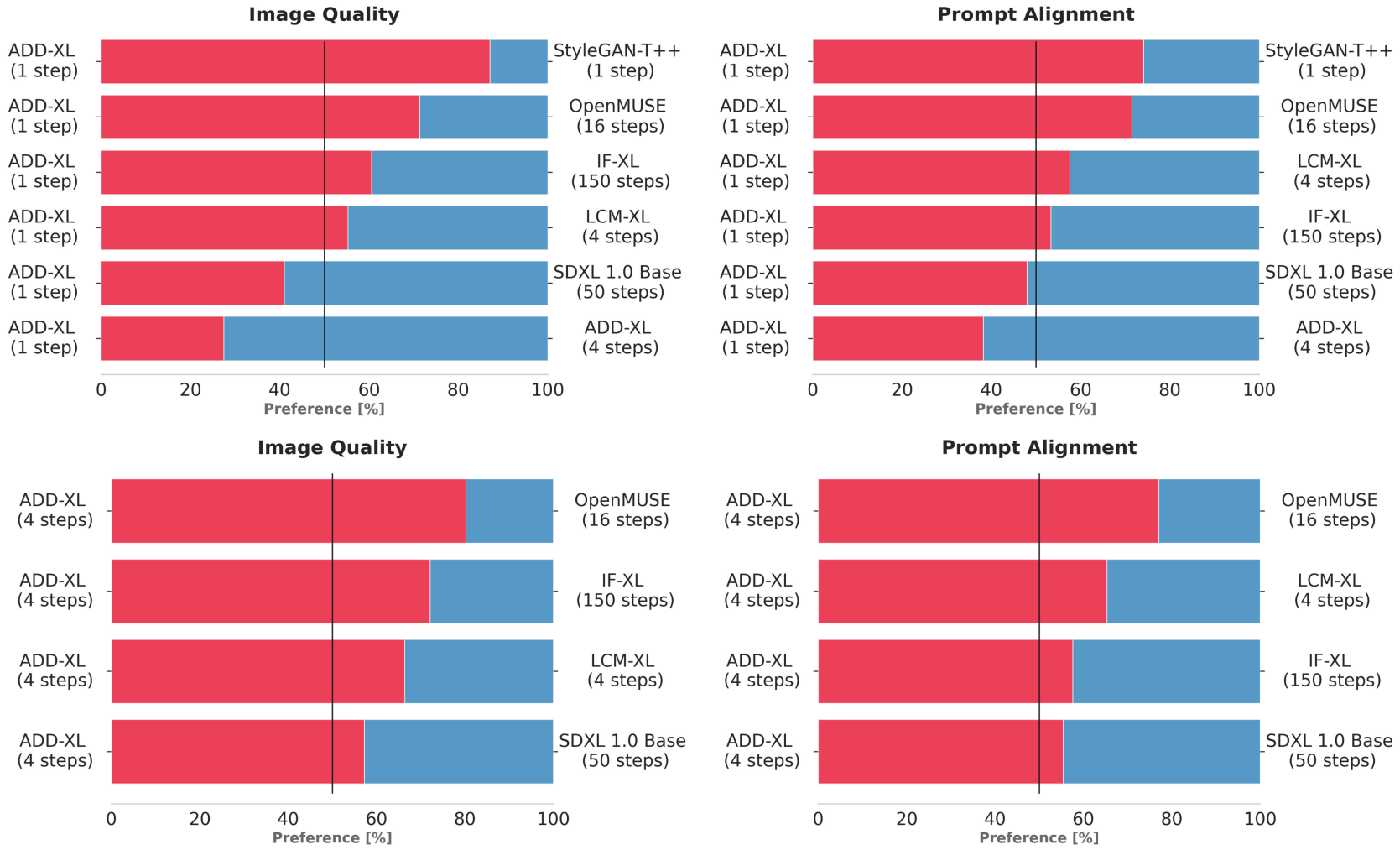


学习Stylegan-t，设计判别器的结构和损失

化简后和2D-3D蒸馏的SDS loss基本一样

Sauer et al. Adversarial Diffusion Distillation.

对抗性扩散蒸馏：结果



真实用户偏好结果：1-4步均优于StyleGAN-T++、LCM-XL、SDXL

在4步的情况下，ADD-XL 甚至超越教师模型 SDXL-Base

Sauer et al. Adversarial Diffusion Distillation.

对抗性扩散蒸馏：1步生成结果

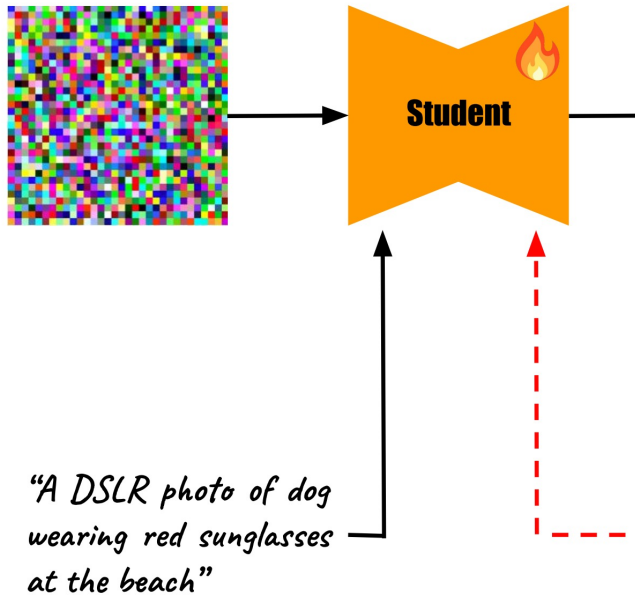
Sauer et al. Adversarial Diffusion Distillation.

对抗性扩散蒸馏：2-4步生成结果

局限：不能超过4步

Sauer et al. Adversarial Diffusion Distillation.

SwiftBrush：VSD替换SDS、不需要对抗损失



Algorithm 1 SwiftBrush Distillation

Require: a pretrained text-to-image teacher ϵ_ψ, a LoRA teacher ϵ_ϕ, a student model f_θ, two learning rates η_1 and η_2, a weighting function ω, a prompts dataset Y, the maximum number of time steps T and the noise schedule $\{(\alpha_t, \sigma_t)\}_{t=1}^{T}$ of the teacher model
Initialize: $\phi \leftarrow \psi, \theta \leftarrow \psi$
while not converged **do**
  Sample input noise $z \sim \mathcal{N}(0, I)$
  Sample text caption input $y \sim Y$
  Compute student output $\hat{x}_0 = f_\theta(z, y)$
  Sample timestep $t \sim \mathcal{U}(0.02T, 0.98T)$
  Sample added noise $\epsilon \sim \mathcal{N}(0, I)$
  Compute noisy sample $\hat{x}_t = \alpha_t \hat{x}_0 + \sigma_t \epsilon$
  $\theta \leftarrow \theta - \eta_1 \left[\omega(t)\left(\epsilon_\psi(\hat{x}_t, t, y) - \epsilon_\phi(\hat{x}_t, t, y)\right)\frac{\partial \hat{x}_0}{\partial \theta}\right]$
  Sample timestep $t' \sim \mathcal{U}(0, T)$
  Sample added noise $\epsilon' \sim \mathcal{N}(0, I)$
  Compute noisy sample $\hat{x}_{t'} = \alpha_{t'} \hat{x}_0 + \sigma_{t'} \epsilon'$
  $\phi \leftarrow \phi - \eta_2 \nabla_\phi \|\epsilon_\phi(\hat{x}_{t'}, t', y) - \epsilon'\|^2$
end while
return trained student model f_θ

$$\min_{\epsilon_\phi} \mathbb{E}_{t,c,\epsilon} \|\epsilon_\phi(x_t, t, y, c) - \epsilon\|_2^2$$

$\mathcal{L}_{\text{diff}}$

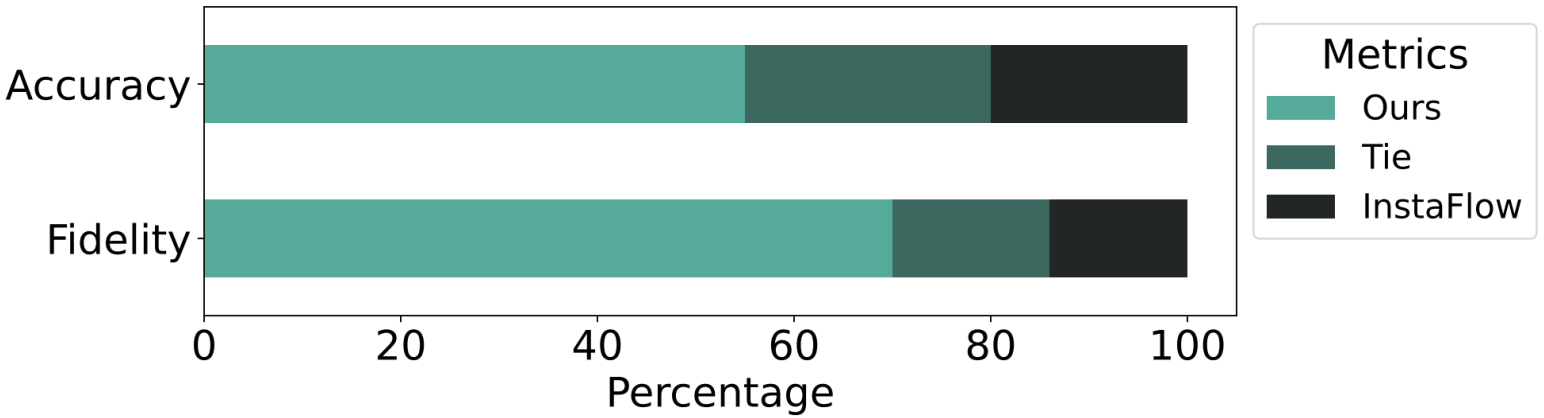
免为预测teacher的输出

$\mathcal{L}_{\text{VSD}}$

$$\nabla_\theta \mathcal{L}_{VSD} = \mathbb{E}_{t,\epsilon,c}[w(t)(\epsilon_\psi(x_t, t, y) - \epsilon_\phi(x_t, t, y, c))\frac{\partial g(\theta, c)}{\partial \theta}]$$

Nguyen & Tran. SwiftBrush : One-Step Text-to-Image Diffusion Model with Variational Score Distillation.

SwiftBrush：结果



人工评测

综合来看可能不如**ADD**，侧面反映了对抗的重要性

Method	Steps	FID-30K ↓	CLIP-30K ↑
Guided Distillation†	1	37.3	0.27
LCM†	1	35.56	0.24
Instaflow	1	**13.10**†	0.28§
BOOT‡	1	48.20	0.26
Ours	1	16.67	**0.29**
SD 2.1*	25	13.45	0.23
SD 2.1*	1	202.14	0.06

COCO 2014上的量化指标

Nguyen & Tran. SwiftBrush : One-Step Text-to-Image Diffusion Model with Variational Score Distillation.

SwiftBrush： 1步生成结果

Nguyen & Tran. SwiftBrush : One-Step Text-to-Image Diffusion Model with Variational Score Distillation.

报告总结

- 扩散模型的理论和方法基础
  - 隐变量模型 **vs.** 评分函数估计
  - 无限时间步的扩展
  - 网络结构：**U-Net vs. U-ViT**
- 扩散模型的方法研究：更可靠、普适、高效
  - 基础改进
    - 对预训练分数预测模型的完善、校准、增强
    - 面向更高分辨率、多模态
  - 生成加速
    - 设计**ODE Solver**
    - 模型蒸馏、一致性模型

感谢各位专家！敬请批评指正！

邮箱： zhijied@sjtu.edu.cn

主页： https://thudzj.github.io/

Github