

# **Diffusion Models: A Comprehensive Survey of Methods and Applications**

LING YANG\*, Peking University, China

ZHILONG ZHANG\*, Peking University, China

YANG SONG, OpenAI, USA

SHENDA HONG, Peking University, China

RUNSHENG XU, University of California, Los Angeles, USA

YUE ZHAO, Carnegie Mellon University, USA

WENTAO ZHANG, Mila - Québec AI Institute, HEC Montréal, Canada

BIN CUI, Peking University, China

MING-HSUAN YANG, University of California at Merced, USA

警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址：[https://github.com/binary-husky/gpt\\_academic/](https://github.com/binary-husky/gpt_academic/)。项目在线体验地址：<https://chatpaper.org>。当前大语言模型:gpt-3.5-turbo，当前语言模型温度设定: 1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

扩散模型作为一种强大的新型深度生成模型家族，在许多应用中创下了突破性的表现，包括图像合成、视频生成和分子设计。在本调查中，我们对扩散模型的快速扩展研究进行了概述，将研究分为三个关键领域：高效采样、改进的似然估计和处理具有特殊结构的数据。我们还讨论了将扩散模型与其他生成模型结合以获得增强结果的潜力。此外，我们还回顾了扩散模型在计算机视觉、自然语言处理、时间数据建模以及其他科学学科的跨学科应用中的广泛应用。本调查旨在提供扩散模型现状的上下文化、深入的了解，明确关注重点

---

\*These authors contributed equally.

---

Authors' addresses: Ling Yang, Peking University, China, yangling0818@163.com; Zhilong Zhang, Peking University, China, zhilong.zhang@bjmu.edu.cn; Yang Song, OpenAI, USA, songyang@openai.com; Shenda Hong, Peking University, China, hongshenda@pku.edu.cn; Runsheng Xu, University of California, Los Angeles, USA, rxx3386@ucla.edu; Yue Zhao, Carnegie Mellon University, USA, zhaoy@cmu.edu; Wentao Zhang, Mila - Québec AI Institute, HEC Montréal, Canada, wentao.zhang@mila.quebec; Bin Cui, Peking University, China, bin.cui@pku.edu.cn; Ming-Hsuan Yang, University of California at Merced, USA, mhyang@ucmerced.edu.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.  
© 2022 Association for Computing Machinery.

Manuscript submitted to ACM

领域，并指出潜在的进一步研究领域。Github链接：<https://github.com/YangLing0818/Diffusion-Models-Papers-Survey-Taxonomy>。

CCS Concepts: • Computing methodologies → Computer vision tasks; Natural language generation; Machine learning approaches.

Additional Key Words and Phrases: Generative Models, Diffusion Models, Score-Based Generative Models, Stochastic Differential Equations

**ACM Reference Format:**

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. 1, 1 (July 2022), 55 pages. <https://doi.org/XXXXXX.XXXXXXX>

## 目录

摘要	1
目录	2
1    Introduction	4
2    Foundations of Diffusion Models	6
2.1    Denoising Diffusion Probabilistic Models (DDPMs)	6
2.2    Score-Based Generative Models (SGMs)	9
2.3    Stochastic Differential Equations (Score SDEs)	10
3    Diffusion Models with Efficient Sampling	11
3.1    Learning-Free Sampling	12
3.1.1    SDE Solvers	13
3.1.2    ODE solvers	14
3.2    Learning-Based Sampling	15
3.2.1    Optimized Discretization	15
3.2.2    Truncated Diffusion	15
3.2.3    Knowledge Distillation	15
4    Diffusion Models with Improved Likelihood	16
4.1    Noise Schedule Optimization	16
4.2    Reverse Variance Learning	17
4.3    Exact Likelihood Computation	17
5    Diffusion Models for Data with Special Structures	19
5.1    Discrete Data	19
5.2    Data with Invariant Structures	19

Diffusion Models: A Comprehensive Survey of Methods and Applications	3
5.3     Data with Manifold Structures	20
5.3.1     Known Manifolds	20
5.3.2     Learned Manifolds	20
6     Connections with Other Generative Models	21
6.1     Variational Autoencoders and Connections with Diffusion Models	22
6.2     Generative Adversarial Networks and Connections with Diffusion Models	23
6.3     Normalizing Flows and Connections with Diffusion Models	24
6.4     Autoregressive Models and Connections with Diffusion Models	24
6.5     Energy-based Models and Connections with Diffusion Models	25
7     Applications of Diffusion Models	26
7.1     Unconditional and Conditional Diffusion Models	26
7.1.1     Conditioning Mechanisms in Diffusion Models	26
7.1.2     Condition Diffusion on Labels and Classifiers	27
7.1.3     Condition Diffusion on Texts, Images, and Semantic Maps	27
7.1.4     Condition Diffusion on Graphs	27
7.2     Computer Vision	27
7.2.1     Image Super Resolution, Inpainting, Restoration, Translation, and Editing	27
7.2.2     Semantic Segmentation	29
7.2.3     Video Generation	29
7.2.4     Point Cloud Completion and Generation	29
7.2.5     Anomaly Detection	30
7.3     Natural Language Generation	30
7.4     Multi-Modal Generation	30
7.4.1     Text-to-Image Generation	30
7.4.2     Scene Graph-to-Image Generation	32
7.4.3     Text-to-3D Generation	34
7.4.4     Text-to-Motion Generation	34
7.4.5     Text-to-Video Generation	34
7.4.6     Text-to-Audio Generation	35
7.5     Temporal Data Modeling	35
7.5.1     Time Series Imputation	35
7.5.2     Time Series Forecasting	37
7.5.3     Waveform Signal Processing	37
7.6     Robust Learning	38

7.7	Interdisciplinary Applications	38
7.7.1	Drug Design and Life Science	38
7.7.2	Material Design	39
7.7.3	Medical Image Reconstruction	39
8	Future Directions	39
	重新审视假设	40
	理论理解	40
	潜在表示	40
	AIGC和扩散基础模型	40
9	Conclusion	40
	References	40

## 1 INTRODUCTION

扩散模型 [97, 247, 252, 257] 已经成为最先进的深度生成模型家族。它们打破了生成对抗网络 (GAN) [77] 在图像合成 [51, 97, 252, 257] 这一具有挑战性的任务中的长期统治地位，并且还在各个领域显示出潜力，涵盖范围广泛，从计算机视觉 [2, 10, 19, 22, 98, 100, 125, 128, 149, 168, 180, 198, 230, 232, 280, 309, 310, 329, 337]，自然语言处理 [6, 103, 153, 237, 316]，时序数据建模 [1, 31, 138, 222, 262, 300]，多模态建模 [7, 216, 228, 231, 335]，鲁棒机器学习 [17, 26, 124, 274, 312]，到交叉学科应用领域，如计算化学 [3, 101, 114, 145, 147, 170, 293] 和医学图像重建 [24, 38–40, 44, 176, 203, 256, 294]。

已经提出了许多方法来改进扩散模型，无论是通过增强实证性能 [188, 249, 253] 还是从理论角度扩展模型的能力 [161, 162, 251, 257, 324]。在过去的两年里，扩散模型的研究成果显著增加，使得新研究人员越来越难以跟上领域中的最新发展。此外，大量的研究工作可能掩盖了主要趋势，阻碍了进一步的研究进展。本综述旨在通过全面概述扩散模型研究的现状，对各种方法进行分类，并突出关键进展，从而解决这些问题。我们希望本综述能够为初入该领域的研究人员提供有用的参考，同时为有经验的研究人员提供更广泛的视角。

在本文中，我们首先解释了扩散模型的基础 (Section 2)，简明但自成体系地介绍了三种主要形式的模型：去噪扩散概率模型 (DDPMs) [97, 247]，基于分数的生成模型 (SGMs) [252, 253] 和随机微分方程 (Score SDEs) [121, 251, 257]。所有这些方法的关键是在随机噪声随着时间逐渐增强的过程中逐步扰动数据（称为“扩散”过程），然后逐步去除噪声以生成新的数据样本。我们阐明了它们如何在扩散的同一原理下工作，并解释了这三个模型如何相互连接并彼此归纳。接下来，我们提出了最近研究的一个分类体系，将扩散模型领域分为三个主要领域：高效采样 (第3节)，改进似然估计 (第4节)，以及处理具有特殊结构数据的方法 (第5节)，例如关系数据、具有排列/旋转不变性的数据以及驻留在流形上的数据。我们进一步

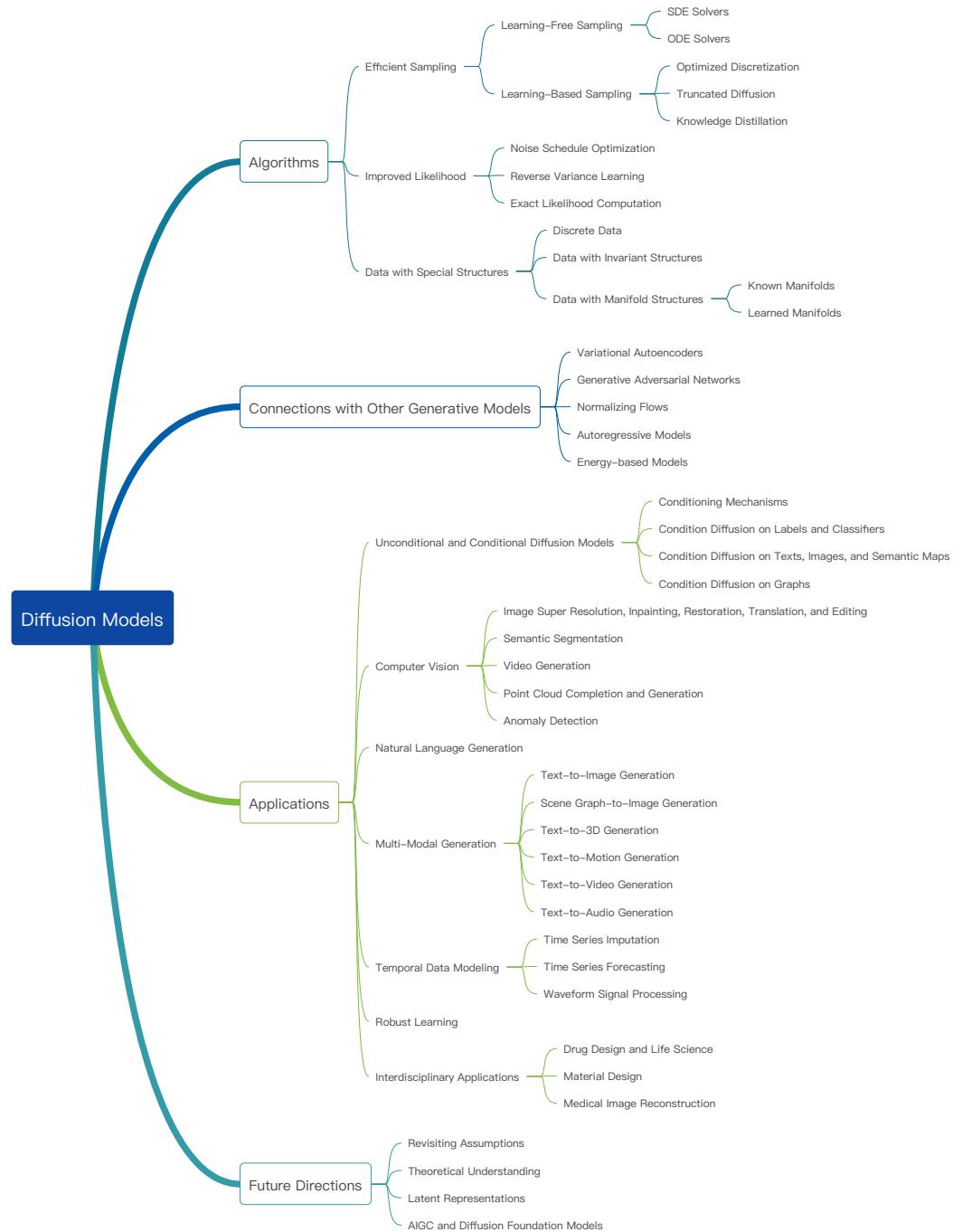


Fig. 1. 扩散模型变体的分类（参见第3节到第5节），与其他生成模型的联系（参见第6节），扩散模型的应用（参见第7节）和未来发展方向（参见第8节）。

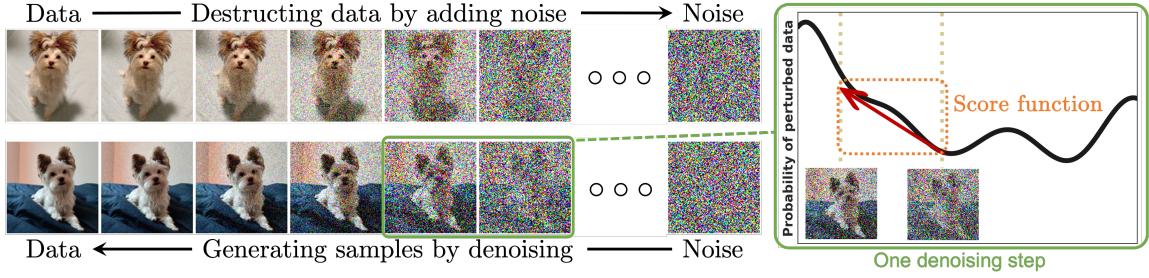


Fig. 2. 扩散模型通过添加噪声，然后通过逆过程从噪声中生成新的数据来平滑扰动数据。逆过程中的每个去噪步骤通常需要估计评分函数（见右侧的示意图），评分函数是一个梯度，指向具有更高可能性和较少噪声的数据方向。

通过将每个类别细分为更详细的子类别来对这些模型进行研究，如图Fig. 1所示。此外，我们还讨论了扩散模型与其他深度生成模型（第6节）之间的联系，包括变分自动编码器（VAEs）[135, 225]、生成对抗网络（GANs）[77]、归一化流[53, 55, 199, 227]、自回归模型[271]和能量-based模型（EBMs）[144, 255]。通过将这些模型与扩散模型相结合，研究人员有可能实现更强大的性能。

紧接着，我们的调查回顾了扩散模型在现有研究中应用到的六个主要应用领域（第7节）：计算机视觉、自然语言处理、时间数据建模、多模态学习、鲁棒学习和跨学科应用。针对每个任务，我们提供一个定义，描述了扩散模型如何用于解决该任务，并总结相关的先前工作。我们通过对这个令人兴奋的新研究领域未来可能的发展方向进行了展望，并总结了我们的论文（第8节，第9节）。

## 2 FOUNDATIONS OF DIFFUSION MODELS

扩散模型是一类概率生成模型，通过注入噪声来逐步破坏数据，然后学习反向过程以进行样本生成。我们在图2中展示了扩散模型的直观理解。目前，对扩散模型的研究主要基于三种主要的公式：去噪扩散概率模型（DDPMs）[97, 188, 247]，基于得分的生成模型（SGMs）[252, 253]和随机微分方程（得分SDEs）[251, 257]。我们在本节中对这三种公式进行了自包含的介绍，并讨论了它们之间的相互联系。

### 2.1 Denoising Diffusion Probabilistic Models (DDPMs)

一种去噪扩散概率模型 (DDPM) [97, 247] 利用了两个马尔可夫链：一个前向链将数据转化为噪声，另一个反向链将噪声转化回数据。前者通常是手工设计的，旨在将任何数据分布转化为简单的先验分布（如标准高斯分布），而后的马尔可夫链通过学习由深度神经网络参数化的过渡核心反转前者。随后，通过首先从先验分布中采样一个随机向量，然后通过反向马尔可夫链进行祖先采样来生成新的数据点 [137]。

形式上，给定一个数据分布  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ ，前向马尔可夫过程生成一系列随机变量  $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_T$ ，其过渡核心为  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$ 。利用概率的链式法则和马尔可夫性质，我们可以将在给定  $\mathbf{x}_0$  的条件下， $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_T$  的联合分布  $q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0)$  分解为

$$q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}). \quad (1)$$

在深度动力学生成模型中，我们手工设计了过渡核函数  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$  来逐步将数据分布  $q(\mathbf{x}_0)$  转化为可处理的先验分布。过渡核的一个典型设计是高斯扰动，而过渡核的最常见选择是

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (2)$$

其中  $\beta_t \in (0, 1)$  是在模型训练之前选择的超参数。我们在这里使用这个核函数仅仅是为了简化我们的讨论，虽然其他类型的核函数也适用。正如Sohl-Dickstein等人(2015) [247]所观察到的，这个高斯转移核使得我们可以边缘化Eq.(1)中的联合分布，从而获得所有  $t \in \{0, 1, \dots, T\}$  的  $q(\mathbf{x}_t | \mathbf{x}_0)$  的解析形式。具体而言，定义  $\alpha_t := 1 - \beta_t$  和  $\bar{\alpha}_t := \prod_{s=0}^t \alpha_s$ ，那么我们有

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}). \quad (3)$$

给定  $\mathbf{x}_0$ ，我们可以通过采样一个服从高斯分布的向量  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ，然后应用变换来获得  $\mathbf{x}_t$  的样本。

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}. \quad (4)$$

当  $\bar{\alpha}_T \approx 0$ ， $\mathbf{x}_T$  的分布几乎服从高斯分布，因此我们有  $q(\mathbf{x}_T) := \int q(\mathbf{x}_T | \mathbf{x}_0) q(\mathbf{x}_0) d\mathbf{x}_0 \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ 。

直观地说，这个前向过程会逐渐向数据注入噪声，直到所有结构都丢失。为了生成新的数据样本，DDPMs 首先从先验分布中生成一个无结构的噪声向量（通常很容易获得），然后通过在逆时间方向上运行可学习的马尔科夫链逐渐去除其中的噪声。具体而言，逆向马尔科夫链由一个先验分布  $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$  和一个可学习的转移核  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  参数化。我们选择先验分布  $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ ，因为前向过程被构建成使得  $q(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ 。可学习的转移核  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  的形式为

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \quad (5)$$

其中， $\theta$ 代表模型参数，均值 $\mu_\theta(\mathbf{x}_t, t)$ 和方差 $\Sigma_\theta(\mathbf{x}_t, t)$ 由深度神经网络参数化。有了这个逆马尔科夫链，我们可以通过首先从噪声向量 $\mathbf{x}_T \sim p(\mathbf{x}_T)$ 中进行采样，然后从可学习的转移核 $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ 中迭代地进行采样，直到 $t = 1$ ，生成一组数据样本 $\mathbf{x}_0$ 。

这个采样过程的成功关键在于训练反向马尔科夫链以匹配正向马尔科夫链的实际时间反转。也就是说，我们必须调整参数 $\theta$ ，使得反向马尔科夫链的联合分布 $p_\theta(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T) := p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ 尽可能地接近正向过程的联合分布 $q(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T) := q(\mathbf{x}_0) \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})$ (Eq. (1))。这通过最小化这两者之间的Kullback-Leibler (KL) 散度来实现：

$$\text{KL}(q(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T) || p_\theta(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)) \quad (6)$$

$$\stackrel{(i)}{=} -\mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)} [\log p_\theta(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)] + \text{const} \quad (7)$$

$$\stackrel{(ii)}{=} \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)} \underbrace{\left[ -\log p(\mathbf{x}_T) - \sum_{t=1}^T \log \frac{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right]}_{:= -L_{\text{VLB}}(\mathbf{x}_0)} + \text{const} \quad (8)$$

$$\stackrel{(iii)}{\geq} \mathbb{E} [-\log p_\theta(\mathbf{x}_0)] + \text{const}, \quad (9)$$

这是由KL散度定义 (i)， $q(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)$  和  $p_\theta(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T)$  都是分布的乘积 (ii)，以及Jensen不等式的结果 (iii) 得出的。公式(Eq. (8))中的第一项是数据 $\mathbf{x}_0$ 的变分下界 (VLB)，这是训练概率生成模型的常见目标。我们用“const”表示一个不依赖于模型参数 $\theta$ 的常数，因此不影响优化。DDPM训练的目标是最大化VLB (或等价地，最小化负的VLB)，这个目标特别容易优化，因为它是独立项的求和，可以通过Monte Carlo采样 [186]和随机优化 [258]高效地估计和优化。

Ho等人 (2020) [97]提议对 $L_{\text{VLB}}$ 中的各项进行重新加权以获得更好的样本质量，并注意到了得到的损失函数与噪声条件的评分网络 (NCSNs) 的训练目标具有重要的等价性，后者是一种基于评分的生成模型，详细可以参考Song和Ermon [252]。在[97]中的损失函数形式如下：

$$\mathbb{E}_{t \sim \mathcal{U}[\![1, T]\!], \mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\lambda(t) \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2] \quad (10)$$

其中， $\lambda(t)$ 是一个正的加权函数， $\mathbf{x}_t$ 由Eq.(4)通过 $\mathbf{x}_0$ 和 $\boldsymbol{\epsilon}$ 计算得到， $\mathcal{U}[\![1, T]\!]$ 是在集合 $\{1, 2, \dots, T\}$ 上的均匀分布，而 $\boldsymbol{\epsilon}_\theta$ 是一个具有参数 $\theta$ 的深度神经网络，给定 $\mathbf{x}_t$ 和 $t$ ，它预测噪声向量 $\boldsymbol{\epsilon}$ 。这个目标函数在选择特定的加权函数 $\lambda(t)$ 后可以化简为Eq. (8)，并具有与训练基于得分的生成模型的多个噪声尺度下降噪声得分匹配的损失相同的形式[252]，这是扩散模型的另一种形式，在下一节将讨论。

## 2.2 Score-Based Generative Models (SGMs)

在评分-based生成模型的核心 [252, 253] 是概念的 (*Stein*) 评分（也称为评分或评分函数）[108]。给定概率密度函数  $p(\mathbf{x})$ ，其评分函数被定义为概率密度的对数的梯度  $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ 。与统计学中常用的 *Fisher* 评分  $\nabla_{\theta} \log p_{\theta}(\mathbf{x})$  不同，此处考虑的是与数据  $\mathbf{x}$  相关的 *Stein* 评分，而不是模型参数  $\theta$ 。它是一个指向概率密度函数增长最快方向的向量场。

评分-based生成模型 (SGMs) 的关键思想 [252] 是通过一系列逐渐增强的高斯噪声扰动数据，并通过在噪声水平上训练一个条件于噪声水平的深度神经网络模型（称为噪声条件评分网络，NCSN，见 [252]）来共同估计所有噪声数据分布的评分函数。通过在不断降低的噪声水平上使用基于评分的采样方法，包括 Langevin Monte Carlo [85, 118, 200, 252, 257]，随机微分方程 [117, 257]，普通微分方程 [121, 162, 251, 257, 324] 以及它们的各种组合 [257]，我们可以生成样本。训练和采样在评分-based生成模型的形式化中完全解耦，因此在估计评分函数之后可以使用多种采样技术。

在 Section 2.1 中使用类似的符号，我们将  $q(\mathbf{x}_0)$  定义为数据分布，并且  $0 < \sigma_1 < \sigma_2 < \dots < \sigma_t < \dots < \sigma_T$  是一个噪声水平的序列。SGM 的典型示例涉及使用高斯噪声分布  $q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma_t^2 \mathbf{I})$  将数据点  $\mathbf{x}_0$  扰动为  $\mathbf{x}_t$ 。这产生了一系列噪声数据密度函数  $q(\mathbf{x}_1), q(\mathbf{x}_2), \dots, q(\mathbf{x}_T)$ ，其中  $q(\mathbf{x}_t) := \int q(\mathbf{x}_t) q(\mathbf{x}_0) d\mathbf{x}_0$ 。噪声条件评分网络是一个深度神经网络  $\mathbf{s}_{\theta}(\mathbf{x}, t)$ ，用于估计评分函数  $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t)$ 。学习评分函数（也称为评分估计）有各种已经建立的技术，如评分匹配 [108]，去噪评分匹配 [218, 219, 272] 和分片评分匹配 [254]，因此我们可以直接使用其中的一种方法来训练我们的噪声条件评分网络。例如，使用去噪评分匹配和 Eq. (10) 中的类似符号，训练目标给定为

$$\mathbb{E}_{t \sim \mathcal{U}[[1, T]], \mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \lambda(t) \sigma_t^2 \|\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t) - \mathbf{s}_{\theta}(\mathbf{x}_t, t)\|^2 \right] \quad (11)$$

$$\stackrel{(i)}{=} \mathbb{E}_{t \sim \mathcal{U}[[1, T]], \mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \lambda(t) \sigma_t^2 \|\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t | \mathbf{x}_0) - \mathbf{s}_{\theta}(\mathbf{x}_t, t)\|^2 \right] + \text{const} \quad (12)$$

$$\stackrel{(ii)}{=} \mathbb{E}_{t \sim \mathcal{U}[[1, T]], \mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \lambda(t) \left\| -\frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma_t} - \sigma_t \mathbf{s}_{\theta}(\mathbf{x}_t, t) \right\|^2 \right] + \text{const} \quad (13)$$

$$\stackrel{(iii)}{=} \mathbb{E}_{t \sim \mathcal{U}[[1, T]], \mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(0, \mathbf{I})} [\lambda(t) \|\epsilon + \sigma_t \mathbf{s}_{\theta}(\mathbf{x}_t, t)\|^2] + \text{const}, \quad (14)$$

其中，(i) 是由 [272] 推导得到的，(ii) 是基于假设  $q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma_t^2 \mathbf{I})$  得到的，(iii) 是基于事实  $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$  得到的。同样，我们用  $\lambda(t)$  表示一个正权重函数，“const” 表示不依赖于可训练参数  $\theta$  的常数。比较 Eq. (14) 和 Eq. (10)，可以明显看出 DDPMs 和 SGMs 的训练目标是等价的，一旦我们设置  $\epsilon_{\theta}(\mathbf{x}, t) = -\sigma_t \mathbf{s}_{\theta}(\mathbf{x}, t)$ 。此外，我们还可以推广高阶得分匹配。数据密度的高阶导数提供了关于数据分布的附加局部信息。Meng 等人 [183] 提出了一种广义的去噪得分匹配方法，

可以有效地估计高阶得分函数。该提出的模型可以提高 Langevin 动力学的混合速度，从而提高扩散模型的采样效率。

对于样本生成，SGMs利用迭代方法依次生成来自  $\mathbf{s}_\theta(\mathbf{x}, T), \mathbf{s}_\theta(\mathbf{x}, T-1), \dots, \mathbf{s}_\theta(\mathbf{x}, 0)$  的样本。由于SGMs中训练和推断的解耦，存在许多采样方法，其中一些将在下一节中讨论。在这里我们介绍SGMs的第一种采样方法，称为退火 Langevin 动力学（ALD）[252]。设  $N$  是每个时间步迭代次数， $s_t > 0$  是步长。我们首先用  $\mathbf{x}_T^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  初始化ALD，然后对  $t = T, T-1, \dots, 1$  依次应用 Langevin Monte Carlo。在每个时间步  $0 \leq t < T$ ，我们从  $\mathbf{x}_t^{(0)} = \mathbf{x}_{t+1}^{(N)}$  开始，然后按照以下更新规则迭代，其中  $i = 0, 1, \dots, N-1$ ：

$$\begin{aligned}\epsilon^{(i)} &\leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{x}_t^{(i+1)} &\leftarrow \mathbf{x}_t^{(i)} + \frac{1}{2}s_t \mathbf{s}_\theta(\mathbf{x}_t^{(i)}, t) + \sqrt{s_t} \epsilon^{(i)}.\end{aligned}$$

Langevin Monte Carlo理论[200]保证当  $s_t \rightarrow 0$  且  $N \rightarrow \infty$  时， $\mathbf{x}_0^{(N)}$  成为从数据分布  $q(\mathbf{x}_0)$  中得到的有效样本。

### 2.3 Stochastic Differential Equations (Score SDEs)

DDPM和SGM可以进一步推广到无限时间步骤或噪声水平的情况，其中扰动和去噪过程是随机微分方程（SDE）的解。我们称之为Score SDE [257]，因为它利用SDE来进行噪声扰动和样本生成，并且去噪过程需要估计有噪数据分布的得分函数。

得分SDE通过以下随机微分方程（SDE）对数据进行扰动到噪声 [257]：

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \quad (15)$$

其中  $\mathbf{f}(\mathbf{x}, t)$  和  $g(t)$  是随机微分方程(SDE)中的扩散项和漂移项， $\mathbf{w}$  是标准维纳过程（又称布朗运动）。DDPMs和SGMs中的前向过程都是此SDE的离散化。正如Song等人在2020年的研究中所示[257]，对于DDPMs，相应的SDE为：

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}dt + \sqrt{\beta(t)}d\mathbf{w} \quad (16)$$

在  $\beta(\frac{t}{T}) = T\beta_t$ ，其中  $T$  趋向无穷大；而对于 SGMs，相应的 SDE 给出如下：

$$d\mathbf{x} = \sqrt{\frac{d[\sigma(t)^2]}{dt}}d\mathbf{w}, \quad (17)$$

当  $T$  趋向于无穷大时，其中  $\sigma(\frac{t}{T}) = \sigma_t$ 。在这里，我们使用  $q_t(\mathbf{x})$  来表示正向过程中  $\mathbf{x}_t$  的分布。

重要的是，对于形式为Eq. (15)的任何扩散过程，Anderson [4]表明，可以通过解下述逆时间随机微分方程来进行反演：

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x})] dt + g(t) d\bar{\mathbf{w}} \quad (18)$$

其中 $\bar{\mathbf{w}}$ 是标准维纳过程，时间倒流时为负无穷小时间步长 $dt$ 。这个反向 SDE 的解轨道与正向 SDE 的边际密度相同，只是在相反的时间方向上演化[257]。直观地说，反向时间 SDE 的解是将噪声逐渐转化为数据的扩散过程。此外，Song 等人 (2020) [257] 证明了一个普通微分方程的存在，即“概率流动微分方程”，其轨迹具有与反向时间 SDE 相同的边际分布。概率流动微分方程如下：

$$d\mathbf{x} = \left[ \mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x}) \right] dt. \quad (19)$$

反向时间随机微分方程 (reverse-time SDE) 和概率流动常微分方程 (probability flow ODE) 允许从相同的数据分布中进行抽样，因为它们的轨迹具有相同的边际概率。

一旦每个时间步骤 $t$ 的得分函数 $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ ，已知，我们可以解开反向时间随机微分方程 (Eq. (18)) 和概率流动常微分方程 (Eq. (19))，并通过各种数值方法生成样本，例如退火朗之万动力学 [252] (cf., Section 2.2)，数值随机微分方程求解器 [117, 257]，数值常微分方程求解器 [121, 162, 249, 257, 324]，以及预测校正方法 (MCMC和数值常微分方程/随机微分方程求解器的组合) [257]。和SGMs中一样，我们通过参数化一个时变的得分模型 $\mathbf{s}_{\theta}(\mathbf{x}_t, t)$ 来估计得分函数，通过将Eq. (14)中的得分匹配目标广义化到连续时间，得到以下目标函数：

$$\mathbb{E}_{t \sim \mathcal{U}[0, T], \mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \lambda(t) \left\| \mathbf{s}_{\theta}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_{0t}(\mathbf{x}_t | \mathbf{x}_0) \right\|^2 \right], \quad (20)$$

其中， $\mathcal{U}[0, T]$ 表示定义在 $[0, T]$ 上的均匀分布，其余符号遵循Eq. (14)。

对扩散模型的后续研究主要集中在改进这些经典方法 (DDPMs、SGMs和Score SDEs) 的三个主要方向上：更快、更高效的采样方法、更准确的似然函数和密度估计，以及处理具有特殊结构的数据（例如置换不变性、流形结构和离散数据）。我们在接下来的三节中对每个方向进行了广泛的概述 (Sections 3 to 5)。在Table 1中，我们列出了在连续和离散时间设置下，三类扩散模型的更详细分类、相应的文章和年份。

### 3 DIFFUSION MODELS WITH EFFICIENT SAMPLING

从扩散模型生成样本通常需要涉及大量评估步骤的迭代方法。最近的很多工作都集中在加速采样过程的同时提高结果样本的质量。我们把这些高效的采样方法分为两类：一类是不涉及

表 1. 在连续和离散的情况下，列出了三种扩散模型，并附上了相应的文章和年份。

Primary	Secondary	Tertiary	Article	Year	Setting	
Learning-Free Sampling	SDE Solvers	SDE Solvers	Song et al. [257]	2020	Continuous	
			Dockhorn et al. [57]	2021	Continuous	
			Jolicoeur et al. [118]	2021	Continuous	
			Jolicoeur et al. [117]	2021	Continuous	
			Chuang et al. [39]	2022	Continuous	
	ODE Solvers		Song et al. [252]	2019	Continuous	
			Karras et al. [121]	2022	Continuous	
			Liu et al. [156]	2021	Continuous	
			Song et al. [249]	2020	Continuous	
			Zhang et al. [325]	2022	Continuous	
Efficient Sampling	Optimized Discretization	Optimized Discretization	Karras et al. [121]	2022	Continuous	
			Lu et al. [162]	2022	Continuous	
			Zhang et al. [324]	2022	Continuous	
			Watson et al. [277]	2021	Discrete	
			Watson et al. [276]	2021	Discrete	
	Knowledge Distillation		Dockhorn et al. [58]	2021	Continuous	
			Salimans et al. [233]	2021	Discrete	
			Luhman et al. [164]	2021	Discrete	
			Meng et al. [179]	2022	Discrete	
			Lyu et al. [173]	2022	Discrete	
Learning-Based Sampling	Truncated Diffusion	Truncated Diffusion	Zheng et al. [331]	2022	Discrete	
			Nichol et al. [188]	2021	Discrete	
			Kingma et al. [133]	2021	Discrete	
			Bao et al. [8]	2021	Discrete	
			Nichol et al. [188]	2021	Discrete	
	Noise Schedule Optimization		Song et al. [251]	2021	Continuous	
			Huang et al. [105]	2021	Continuous	
			Song et al. [257]	2020	Continuous	
			Lu et al. [161]	2022	Continuous	
			Vahdat et al. [268]	2021	Continuous	
Improved Likelihood	Learned Manifolds	Learned Manifolds	Wehenkel et al. [278]	2021	Discrete	
			Ramesh et al. [216]	2022	Discrete	
			Rombach et al. [228]	2022	Discrete	
			Bortoli et al. [47]	2022	Continuous	
			Huang et al. [104]	2022	Continuous	
	Known Manifolds		Niu et al. [193]	2020	Discrete	
			Jo et al. [115]	2022	Continuous	
			Shi et al. [241]	2022	Continuous	
			Xu et al. [298]	2021	Discrete	
			Meng et al. [178]	2022	Discrete	
Data with Special Structures	Data with Invariant Structures	Data with Invariant Structures	liu et al. [158]	2023	Continuous	
			Sohl et al. [247]	2015	Discrete	
			Austin et al. [6]	2021	Discrete	
			Xie et al. [292]	2022	Discrete	
			Gu et al. [87]	2022	Discrete	
	Discrete Data		Campbell et al. [23]	2022	Continuous	

学习的（无学习采样），另一类是在扩散模型训练完成后，需要额外的学习过程（基于学习的采样）。

### 3.1 Learning-Free Sampling

许多扩散模型的取样器依赖于对Eq. (18)中的逆时间SDE或Eq. (19)中的概率流ODE进行离散化。由于取样成本与离散化时间步数成正比增加，许多研究人员致力于开发减少时间步数并最小化离散化误差的离散化方案。

**3.1.1 SDE Solvers.** DDPM的生成过程可以被看作是反向时间SDE的特定离散化。如Section 2.3所讨论的，DDPM的正向过程离散化了Eq. (16)中的SDE，而其对应的反向SDE表达为：

$$dx = -\frac{1}{2}\beta(t)(x_t - \nabla_{x_t} \log q_t(x_t))dt + \sqrt{\beta(t)}dw \quad (21)$$

Song等人（2020年）[257]表明，由Eq. (5)定义的逆马尔可夫链相当于Eq. (21)的数值SDE求解器。

噪声条件评分网络（NCSNs）[252]和临界阻尼朗之万扩散（CLD）[57]都通过从朗之万动力学获得灵感来解决逆时间SDE。特别地，NCSNs利用退火朗之万动力学（ALD，cf., Section 2.2）迭代生成数据，同时平滑降低噪声水平，直到生成的数据分布收敛到原始数据分布。虽然ALD的采样轨迹不是逆时间SDE的精确解，但它们具有正确的边缘分布，因此在朗之万动力学在每个噪声水平上收敛到平衡态的假设下，可以产生正确的样本。ALD方法通过一致退火采样（CAS）[118]进一步改进，这是一种具有更好时间步长缩放和添加噪声的基于评分的MCMC方法。CLD受统计力学的启发，提出了扩展SDE，其中包含一个辅助速度项，类似于欠阻尼朗之万扩散。为了获得扩展SDE的时间反演，CLD只需要学习给定数据的条件速度分布的评分函数，这可以说比直接学习数据的得分函数更容易。据报道，添加的速度项可以提高采样速度和质量。

在[257]中提出的逆扩散方法以与正向扩散方法相同的方式离散化了逆时间SDE。对于正向SDE的任何一步离散化，可以写出如下的一般形式：

$$x_{i+1} = x_i + f_i(x_i) + g_i z_i, \quad i = 0, 1, \dots, N-1 \quad (22)$$

其中  $z_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ， $f_i$  和  $g_i$  是由随机微分方程的漂移/扩散系数和离散化方案所确定的。反向扩散是指类似于正向随机微分方程的方式离散化逆向时间随机微分方程，即

$$x_i = x_{i+1} - f_{i+1}(x_{i+1}) + g_{i+1}^t s_{\theta^*}(x_{i+1}, t_{i+1}) + g_{i+1} z_i \quad i = 0, 1, \dots, N-1 \quad (23)$$

其中， $s_{\theta^*}(x_i, t_i)$  是训练好的噪声条件评分模型。Song等人（2020）[257]证明了逆扩散方法是逆向时间SDE的数值解。这个过程可以应用于任何类型的正向SDE，并且实证结果表明，对于一种特定类型的VP-SDE，该采样器的性能略好于DDPM [257]。

Jolicoeur-Martineau等人（2021）[117]开发了一种自适应步长的SDE求解器，用于更快的生成。步长由高阶SDE求解器的输出与低阶SDE求解器的输出进行比较来控制。在每个时间步中，高阶和低阶求解器分别从先前的样本  $x'_{prev}$  生成新样本  $x'_{high}$  和  $x'_{low}$ 。然后通过比较两个样本之间的差异来调整步长。如果  $x'_{high}$  和  $x'_{low}$  相似，算法将返回  $x'_{high}$  并增加步长。 $x'_{high}$  和  $x'_{low}$  之间的

相似性是通过以下方法测量的：

$$E_q = \left\| \frac{\mathbf{x}'_{\text{low}} - \mathbf{x}'_{\text{high}}}{\delta(\mathbf{x}', \mathbf{x}'_{\text{prev}})} \right\|^2 \quad (24)$$

其中， $\delta(\mathbf{x}'_{\text{low}}, \mathbf{x}'_{\text{prev}}) := \max(\epsilon_{abs}, \epsilon_{rel} \max(|\mathbf{x}'_{\text{low}}|, |\mathbf{x}'_{\text{prev}}|))$ ， $\epsilon_{abs}$  和  $\epsilon_{rel}$  分别为绝对容差和相对容差。

[257] 提出的预测-校正方法通过将数值随机微分方程 (SDE) 求解器 (“预测器”) 与迭代的马尔科夫链蒙特卡洛 (MCMC) 方法 (“校正器”) 结合起来，来解决反向SDE。在每个时间步骤中，预测-校正方法首先使用数值SDE求解器生成粗糙样本，然后使用基于评分的MCMC方法来修正样本的边缘分布。所得到的样本在时间边缘上与反向时间SDE的解轨迹具有相同的分布，即它们在所有时间步骤上的分布是等价的。实证结果表明，基于 Langevin 蒙特卡洛的校正器比使用额外的没有校正器的预测器更高效[257]。Karras 等人 (2022) [121]在[257]中进一步改进了 Langevin 动力学校正器，通过引入类似于“搅拌”步骤的 Langevin 过程来添加和移除噪声，在 CIFAR-10 [140] 和 ImageNet-64 [49] 等数据集上实现了新的样本质量最优。

**3.1.2 ODE solvers.** 大量关于更快扩散采样器的研究是基于解决在Section 2.3中介绍的概率流ODE (Eq. (19))。与SDE求解器相反，ODE求解器的轨迹是确定性的，因此不受随机波动的影响。这些确定性ODE求解器通常比其随机对应物快得多，但样本质量稍差。

去噪扩散隐式模型 (DDIM) [249]是加速扩散模型采样的早期工作之一。最初的动机是将原始的DDPM扩展到非马尔可夫案例，使用以下马尔可夫链。

$$q(\mathbf{x}_1, \dots, \mathbf{x}_T \mid \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0) \quad (25)$$

$$q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1} \mid \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2 \mathbf{I}) \quad (26)$$

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) := \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0}{\sqrt{1 - \bar{\alpha}_t}} \quad (27)$$

该公式以DDPM和DDIM作为特例，其中DDPM对应于设置 $\sigma_t^2 = \frac{\hat{\beta}_{t-1}}{\hat{\beta}_t} \beta_t$ ，而DDIM对应于设置 $\sigma_t^2 = 0$ 。DDIM学习了一个马尔可夫链来反转这个非马尔可夫的扰动过程，在 $\sigma_t^2 = 0$ 时完全是确定性的。[121, 162, 233, 249]观察到DDIM采样过程相当于概率流ODE的一种特殊离散化方案。在对单样本数据集上分析了DDIM后，广义去噪扩散隐式模型(gDDIM) [325]提出了一种修改过的得分网络参数化方法，使得在更一般的扩散过程中实现确定性采样，例如Critically-Damped Langevin Diffusion (CLD) [57]中的扩散过程。PNDM [156]提出了一种伪数值方法，用于在 $\mathcal{R}^N$ 上生成特定流形上的样本。它使用具有非线性转移部分的数值求解器来求解流形上的微分方程，然后生成样本，其中包含DDIM作为特例。

通过广泛的实验研究, Karras等人(2022) [121]显示了Heun的二阶方法[5]在样本质量和采样速度之间提供了出色的平衡。高阶求解器在每个时间步需要额外评估一次学习到的得分函数, 以减小离散化误差。与使用更少采样步骤的Euler方法相比, Heun方法可以生成具有相当甚至更好质量的样本。

扩散指数积分采样器[324]和DPM-solver[162]利用概率流ODE的半线性结构, 开发了比通用龙格-库塔方法更高效的定制ODE求解器。具体而言, 概率流ODE的线性部分可以进行解析计算, 而非线性部分可以使用与ODE求解器领域中的指数积分器类似的技术进行求解。这些方法包含了DDIM作为一阶近似。然而, 它们还允许使用更高阶的积分器, 在仅需10到20次迭代的情况下生成高质量的样本, 远远少于通常需要的数百次扩散模型迭代次数。

### 3.2 Learning-Based Sampling

基于学习的采样是扩散模型的另一种高效方法。通过使用部分步骤或训练逆过程的采样器, 该方法以更快的采样速度为代价, 稍微降低了样本质量。与使用手工步骤的无学习方法不同, 基于学习的采样通常涉及通过优化特定的学习目标来选择步骤。

**3.2.1 Optimized Discretization.** 给定一个预训练的扩散模型, Watson等人 (2021年) [277]提出了一种策略, 通过选择最佳的 $K$ 个时间步骤来最大化DDPMs的训练目标, 以找到最优的离散化方案。这种方法的关键在于观察到DDPM目标可以分解成若干个单独的项的和, 使其非常适合动态规划。然而, 众所周知, 用于DDPM训练的变分下界与样本质量没有直接的相关性[265]。一项随后的工作, 称为可微分扩散采样搜索[276], 通过直接优化称为内核内部距离 (Kernel Inception Distance, 简称KID) 的常用样本质量指标来解决这个问题[16]。这种优化在重参数化[135, 225]和梯度重生材料化的帮助下是可行的。基于截断泰勒方法, Dockhorn等人 (2022年) [58]推导出了一个二阶求解器, 通过在一阶记分网络的顶部训练一个额外的头来加速合成。

**3.2.2 Truncated Diffusion.** 一个可以改进采样速度的方法是截断前向和反向扩散过程[173, 331]。其关键思想是在前向扩散过程刚开始几步之后就停止, 然后用非高斯分布开始反向去噪过程。可以通过从预训练的生成模型 (如变分自编码器[135, 225]或生成对抗网络[77]) 中扩散样本来高效地获得此分布的样本。

**3.2.3 Knowledge Distillation.** 使用知识蒸馏方法可以显著提高扩散模型的采样速度[164, 179, 233]。具体而言, 在Progressive Distillation [233]中, 作者提出将完整的采样过程提炼为一个需要步骤减半的更快的采样器。通过将新采样器参数化为深度神经网络, 作者能够训练采样器以匹配DDIM采样过程的输入和输出。反复执行此过程可以进一步减少采样步骤, 尽管较少的步骤可能导致样本质量降低。为解决这个问题, 作者建议了扩散模型的新参数化方法和目标函数的新加权方案。

## 4 DIFFUSION MODELS WITH IMPROVED LIKELIHOOD

正如在Section 2.1中讨论的那样，扩散模型的训练目标是对数似然的（负）变分下界(VLB)。然而，在许多情况下，这个下界可能不是很紧凑[133]，导致扩散模型的对数似然可能是次优的。在本节中，我们调查了最近关于扩散模型的极大似然估计的研究工作。我们重点关注三种类型的方法：噪声调度优化、逆方差学习和精确对数似然评估。

### 4.1 Noise Schedule Optimization

在传统扩散模型的建模中，前向过程中的噪声安排是手工设计的，没有可训练的参数。通过同时优化扩散模型的前向噪声安排和其他参数，可以进一步最大化变分下界 (VLB)，以实现更高的对数似然值[133, 188]。

iDDPM的工作[188]表明，某种余弦噪声安排可以提高对数似然值。具体来说，他们的工作中的余弦噪声安排采用以下形式：

$$\bar{\alpha}_t = \frac{h(t)}{h(0)}, \quad h(t) = \cos\left(\frac{t/T + m}{1+m} \cdot \frac{\pi}{2}\right)^2 \quad (28)$$

其中 $\bar{\alpha}_t$ 和 $\beta_t$ 在公式2和3中有定义，而 $m$ 是一个超参数，用于控制 $t = 0$ 时的噪声尺度。他们还提出了在对数域中在 $\beta_t$ 和 $1 - \bar{\alpha}_t$ 之间进行反方差参数化的插值。

在变分扩散模型(VDMs)中[133]，作者提出了通过联合训练噪声调度和其他扩散模型参数来最大化VLB，从而改善连续时间扩散模型的似然性。他们使用一个单调神经网络 $\gamma_\eta(t)$ 对噪声调度进行参数化，并根据 $\sigma_t^2 = \text{sigmoid}(\gamma_\eta(t))$ 、 $q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\bar{\alpha}_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$ 和 $\bar{\alpha}_t = \sqrt{(1 - \sigma_t^2)}$ 构建前向扰动过程。此外，作者证明数据点 $\mathbf{x}$ 的VLB可以简化为只依赖于信噪比 $R(t) := \frac{\bar{\alpha}_t^2}{\sigma_t^2}$ 的形式。特别地， $L_{VLB}$ 可以分解为

$$L_{VLB} = -\mathbb{E}_{\mathbf{x}_0} \text{KL}(q(\mathbf{x}_T | \mathbf{x}_0) || p(\mathbf{x}_T)) + \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1} \log p(\mathbf{x}_0 | \mathbf{x}_1) - L_D, \quad (29)$$

第一项和第二项可以直接进行优化，类似于训练变分自编码器。第三项可以进一步简化如下：

$$L_D = \frac{1}{2} \mathbb{E}_{\mathbf{x}_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \int_{R_{\min}}^{R_{\max}} \|\mathbf{x}_0 - \tilde{\mathbf{x}}_\theta(\mathbf{x}_v, v)\|_2^2 dv, \quad (30)$$

其中 $R_{\max} = R(1)$ ， $R_{\min} = R(T)$ ， $\mathbf{x}_v = \bar{\alpha}_v \mathbf{x}_0 + \sigma_v \epsilon$ 表示通过向前扰动过程对 $\mathbf{x}_0$ 扩散直到 $t = R^{-1}(v)$ 获得的带噪数据点，而 $\tilde{\mathbf{x}}_\theta$ 则表示扩散模型预测的无噪数据点。结果是，只要噪声计划在 $R_{\min}$ 和 $R_{\max}$ 处具有相同的值，噪声计划不会影响变分下界 (VB-ELBO)，而只会影响 VB-ELBO 的 Monte Carlo 估计器的方差。

## 4.2 Reverse Variance Learning

扩散模型的经典形式假设反向马尔可夫链中的高斯转移核具有固定的方差参数。回想一下，在Eq. (5)中我们将反向核函数表示为 $q_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$ ，但通常将反向方差 $\Sigma_\theta(\mathbf{x}_t, t)$ 固定为 $\beta_t \mathbf{I}$ 。许多方法建议训练反向方差，以进一步最大化VLB和对数似然值。

在iDDPM中[188]，Nichol和Dhariwal提出通过参数化反向方差并使用混合目标来学习它们，从而实现更高的对数似然值和更快的采样速度，同时不损失样本质量。特别地，他们将Eq. (5)中的反向方差参数化为：

$$\Sigma_\theta(\mathbf{x}_t, t) = \exp(\theta \cdot \log \beta_t + (1 - \theta) \cdot \log \tilde{\beta}_t), \quad (31)$$

其中， $\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t$ ，且 $\theta$ 被联合训练以最大化VLB。这种简单的参数化避免了估计更复杂形式的 $\Sigma_\theta(\mathbf{x}_t, t)$ 的不稳定性，并据报道可以提高似然值。

Analytic-DPM [8] 展示了一个引人注目的结果，即最佳反向方差可以从预训练的评分函数中得到，其解析形式如下：

$$\Sigma_\theta(\mathbf{x}_t, t) = \sigma_t^2 + \left( \sqrt{\frac{\bar{\beta}_t}{\alpha_t}} - \sqrt{\bar{\beta}_{t-1} - \sigma_t^2} \right)^2 \cdot \left( 1 - \bar{\beta}_t \mathbb{E}_{q_t(\mathbf{x}_t)} \frac{\|\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t)\|^2}{d} \right) \quad (32)$$

因此，对于一个预先训练的评分模型，我们可以估计它的一阶和二阶矩，以获取最优的逆方差。将它们插入到变分下界（VLB）中可以得到更紧的变分下界和更高的似然值。

## 4.3 Exact Likelihood Computation

在Score SDE [257]的模型中，样本是通过求解以下逆向SDE生成的，在这个SDE中，Eq. (18)中的 $\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t, t)$ 被学习得到的噪音条件分数模型 $\mathbf{s}_\theta(\mathbf{x}_t, t)$ 所替代：

$$d\mathbf{x} = f(\mathbf{x}_t, t) - g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, t) dt + g(t) d\mathbf{w}. \quad (33)$$

我们在这里使用 $p_\theta^{\text{sde}}$ 来表示通过解上述SDE生成的样本的分布。也可以通过将评分模型插入概率流ODE Eq. (19)中来生成数据，得到：

$$\frac{d\mathbf{x}_t}{dt} = \underbrace{f(\mathbf{x}_t, t) - \frac{1}{2} g^2(t) \mathbf{s}_\theta(\mathbf{x}_t, t)}_{:= \hat{f}_\theta(\mathbf{x}_t, t)} \quad (34)$$

同样地，我们使用 $p_\theta^{\text{ode}}$ 表示通过解决这个ODE生成的样本的分布。神经ODE的理论[32]和连续正规化流[81]表明，尽管计算成本很高，但可以精确计算 $p_\theta^{\text{ode}}$ 。对于 $p_\theta^{\text{sde}}$ ，多个并行工

作[105, 161, 251]证明了存在一个有效计算的变分下界，并且我们可以直接训练扩散模型来最大化  $p_\theta^{\text{sde}}$ ，使用的是修改后的扩散损失。

具体来说，Song等人（2021）[251]证明了通过特定的加权函数（似然加权）训练score SDEs所使用的目标在数据上隐含地最大化了  $p_\theta^{\text{sde}}$  的期望值。结果显示

$$\mathbf{D}_{KL}(q_0 \parallel p_\theta^{\text{sde}}) \leq \mathcal{L}(\theta; g(\cdot)^2) + \mathbf{D}_{KL}(q_T \parallel \pi), \quad (35)$$

其中  $\mathcal{L}(\theta; g(\cdot)^2)$  是公式20中具有  $\lambda(t) = g(t)^2$  的得分SDE目标。由于  $\mathbf{D}_{KL}(q_0 \parallel p_\theta^{\text{sde}}) = -\mathbb{E}_{q_0} \log(p_\theta^{\text{sde}}) + \text{const}$ ，且  $\mathbf{D}_{KL}(q_T \parallel \pi)$  是一个常数，使用  $\mathcal{L}(\theta; g(\cdot)^2)$  进行训练等价于最小化数据上的期望负对数似然  $-\mathbb{E}_{q_0} \log(p_\theta^{\text{sde}})$ 。此外，Song et al. (2021) 和 Huang et al. (2021) [105, 251] 提供了  $p_\theta^{\text{sde}}(\mathbf{x})$  的以下界限：

$$-\log p_\theta^{\text{sde}}(\mathbf{x}) \leq \mathcal{L}'(\mathbf{x}), \quad (36)$$

$\mathcal{L}'(\mathbf{x})$  的定义如下：

$$\mathcal{L}'(\mathbf{x}) := \int_0^T \mathbb{E} \left[ \frac{1}{2} \|g(t)\mathbf{s}_\theta(\mathbf{x}_t, t)\|^2 + \nabla \cdot (g(t)^2 \mathbf{s}_\theta(\mathbf{x}_t, t) - f(\mathbf{x}_t, t)) \mid \mathbf{x}_0 = \mathbf{x} \right] dt - \mathbb{E}_{\mathbf{x}_T} [\log p_\theta^{\text{sde}}(\mathbf{x}_T) \mid \mathbf{x}_0 = \mathbf{x}] \quad (37)$$

Eq. (37)的第一部分类似于隐式评分匹配方法[108]，整个限制可以通过蒙特卡洛方法高效估计。

由于概率流ODE是神经ODE或连续归一化流的特例，我们可以使用这些领域中的成熟方法来准确计算  $\log p_\theta^{\text{oode}}$ 。具体而言，我们有

$$\log p_\theta^{\text{oode}}(\mathbf{x}_0) = \log p_T(\mathbf{x}_T) + \int_{t=0}^T \nabla \cdot \tilde{f}_\theta(\mathbf{x}_t, t) dt. \quad (38)$$

可以利用数值常微分方程（ODE）求解器和Skilling-Hutchinson迹估计器[107, 246]来计算上述一维积分。不幸的是，这个公式无法直接优化以最大化数据上的  $p_\theta^{\text{oode}}$ ，因为它需要为每个数据点  $\mathbf{x}_0$  调用昂贵的ODE求解器。为了降低直接最大化上述公式中  $p_\theta^{\text{oode}}$  的成本，Song等人（2021）[251]提出了最大化  $p_\theta^{\text{sde}}$  的变分下界作为最大化  $p_\theta^{\text{oode}}$  的代理的方法，从而产生了一类称为ScoreFlows的扩散模型。

Lu等人（2022）[161]通过提出不仅最小化基本分数匹配损失函数，还包括其更高阶的推广来进一步改进ScoreFlows。他们证明了  $\log p_\theta^{\text{oode}}$  可以用一阶、二阶和三阶分数匹配误差加以限制。基于这一理论结果，作者进一步提出了用于最小化高阶分数匹配损失的高效训练算法，并且报道了在数据上改进的  $p_\theta^{\text{oode}}$ 。

## 5 DIFFUSION MODELS FOR DATA WITH SPECIAL STRUCTURES

尽管扩散模型在图像和音频等数据领域取得了巨大成功，但它们并不一定能够顺利应用于其他模态。许多重要的数据领域具有特殊结构，必须考虑这些结构以使扩散模型能够有效工作。例如，当模型依赖仅在连续数据领域上定义的评分函数，或者数据驻留在低维流形上时，可能会出现困难。为了应对这些挑战，必须以各种方式对扩散模型进行调整。

### 5.1 Discrete Data

大多数扩散模型都针对连续数据域，这是因为在离散数据中使用的高斯噪声扰动不太适合，并且SGMs和Score SDEs所需的评分函数仅在连续数据域上定义。为了解决这个困难，几篇论文（Hoogeboom等，2021[103]；Gu等，2022[87]；Xie等，2022[292]；Austin等，2021[6]）在Sohl-Dickstein等人（2015）[247]的基础上，提出了生成高维离散数据的方法。具体而言，VQ-Diffusion[87]将高斯噪声替换为在离散数据空间上的随机游走，或者随机掩码操作。得到的正向过程的转移核函数形式为。

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathbf{v}^\top(\mathbf{x}_t) \mathbf{Q}_t \mathbf{v}(\mathbf{x}_{t-1}) \quad (39)$$

其中 $\mathbf{v}(\mathbf{x})$ 是一个单热列向量， $\mathbf{Q}_t$ 是懒惰随机行走的转移核。D3PM [6]通过使用吸收态核或离散化高斯核构建正向噪声过程，适应扩散模型中的离散数据。Campbell等人（2022）[23]提出了离散扩散模型的第一个连续时间框架。通过利用连续时间马尔科夫链，他们能够导出效率更高的采样器，优于离散对应物，并对样本分布与真实数据分布之间的误差进行了理论分析。

具体评分匹配（CSM）[178]提出了离散随机变量的得分函数的概括。具体评分定义为关于输入方向变化的概率变化率，它可以看作连续（Stein）得分的有限差分近似。具体评分可以高效地训练和应用于MCMC。

基于随机微积分理论，Liu等人（2023）[158]提出了一个扩散模型的框架，用于在约束和结构化域中生成数据，包括离散数据作为特殊情况。利用随机微积分中的一个基本定理，Doob的h变换，可以通过在逆扩散过程中包含一个特殊力项来约束数据分布在特定区域。他们使用一个基于EM的优化算法对力项进行参数化。此外，通过Girsanov定理，损失函数可以转换为 $L_2$ 损失。

### 5.2 Data with Invariant Structures

许多重要领域的数据具有不变结构。例如，图形具有置换不变性，点云具有平移和旋转不变性。在扩散模型中，通常忽略这些不变性，从而可能导致性能不佳。为了解决这个问题，一些研究[47, 193]提出了赋予扩散模型对数据的不变性处理能力的方法。

Niu等人(2020)[193]首次使用置换等变图神经网络[78, 238, 288], 称为EDP-GNN, 来参数化噪声条件评分模型, 解决了置换不变图生成的问题。GDSS[115]进一步发展了这个思路, 提出了一个连续时间图扩散过程。该过程通过随机微分方程组(SDEs)对节点和边的联合分布进行建模, 使用消息传递操作来保证置换不变性。

类似地, Shi等人(2021)[241]和Xu等人(2022)[298]使扩散模型能够生成对平移和旋转都是不变的分子构象。例如, Xu等人(2022)[298]显示, 由具有不变先验并以等变马尔科夫核演化的马尔科夫链可以诱导出一个不变的边际分布, 该分布可用于强制分子构象中的适当数据不变性。形式上, 设 $\mathcal{T}$ 为旋转或平移操作。假设 $p(\mathbf{x}_T) = p(\mathcal{T}(\mathbf{x}_T))$ ,  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = p_\theta(\mathcal{T}(\mathbf{x}_{t-1}) | \mathcal{T}(\mathbf{x}_t))$ , Xu等人(2022)[298]证明了样本的分布确保对 $\mathcal{T}$ 不变, 即 $p_0(\mathbf{x}) = p_0(\mathcal{T}(\mathbf{x}))$ 。因此, 只要先验和转移核享有相同的不变性, 就可以构建一个生成旋转和平移不变分子构象的扩散模型。

### 5.3 Data with Manifold Structures

机器学习中普遍存在着具有多样结构的数据。正如流形假设(Fefferman等人, 2016年)所述, 自然数据通常在具有较低内在维度的流形上存在。此外, 许多数据领域具有众所周知的流形结构。例如, 气候和地球数据自然地位于球体上, 这是我们的行星形状。许多研究致力于为流形上的数据开发扩散模型。我们根据流形是已知的还是学习的分类, 并在下面介绍一些代表性的作品。

**5.3.1 Known Manifolds.** 最近的研究将Score SDE的表述扩展到多种已知的流形。这种适应性类似于神经ODE的泛化[32]和连续正规化流的泛化[81]到Riemannian流形[160, 175]。为了训练这些模型, 研究人员还将得分匹配和得分函数适应到Riemannian流形上。

Riemannian基于得分的生成模型(RSGM)[47]适用于广泛的流形, 包括球体和环面, 只要它们满足一些温和的条件。RSGM证明了将扩散模型推广到紧致Riemannian流形上是可能的。该模型还提供了在流形上进行扩散逆过程的公式。以内在视角看, RSGM使用测地线随机游走来近似Riemannian流形上的采样过程。该模型通过一个广义去噪得分匹配目标来进行训练。

相比之下, Riemannian扩散模型(RDM)[104]采用变分框架将连续时间扩散模型推广到Riemannian流形上。RDM使用对数似然的变分下界(VLB)作为其损失函数。RDM模型的作者表明, 最大化这个VLB等同于最小化一个Riemannian得分匹配损失。与RSGM不同, RDM采用外在视角, 假设相关的Riemannian流形嵌入在更高维的欧几里得空间中。

**5.3.2 Learned Manifolds.** 根据多样流形假设(Fefferman等人, 2016), 大多数自然数据位于具有显著降维的流形上。因此, 识别这些流形并直接在其上训练扩散模型具有优势, 因为数据维度较低。许多最近的研究建立在这个思想上, 首先使用自编码器将数据压缩到较低维度的流形上, 然后在这个潜空间中训练扩散模型。在这些情况下, 流形是由自编码器隐式定义和通

过重建损失学习的。为了取得成功，设计一个允许自编码器和扩散模型联合训练的损失函数至关重要。

潜在评分生成模型（LSGM）(Vahdat等人, 2021)通过将评分SDE扩散模型与变分自编码器（VAE）(Kingma和Welling, 2013; Rezende等人, 2014)配对，解决了联合训练的问题。在这种配置下，扩散模型负责学习先验分布。LSGM的作者提出了一个联合训练目标，将VAE的证据下界与扩散模型的评分匹配目标结合起来。这导致了一个新的数据对数似然的下界。通过将扩散模型放置在潜空间中，LSGM实现了比传统扩散模型更快的样本生成。此外，LSGM可以通过将离散数据转换为连续的潜变量来处理离散数据。

与联合训练自编码器和扩散模型不同，潜在扩散模型（LDM）(Rombach等人, 2022)分别处理每个组件。首先，训练一个自编码器产生低维潜空间。然后，训练一个扩散模型生成潜变量编码。DALLE-2 (Ramesh等人, 2022)采用类似的策略，通过在CLIP图像嵌入空间上训练一个扩散模型，然后训练一个单独的解码器根据CLIP图像嵌入创建图像。

## 6 CONNECTIONS WITH OTHER GENERATIVE MODELS

在本节中，我们首先介绍了另外五个重要的生成模型类别，并分析了它们的优点和局限性。然后我们介绍了扩散模型与它们的关联，并说明了如何通过融合扩散模型来提升这些生成模型的性能。将扩散模型与其他生成模型整合的算法总结在Table 2中，并在Fig. 3中提供了图解说明。

表 2. 扩散模型被纳入不同的生成模型中。

Model	Article	Year
VAE	Luo et al. [165]	2022
	Hunag et al. [105]	2021
	Vadhat et al. [268]	2021
GAN	Wang et al. [275]	2022
	Xiao et al. [290]	2021
Normalizing Flow	Zhang et al.[323]	2021
	Gong et al. [76]	2021
	kim et al. [129]	2022
Autoregressive Model	Meng et al.[184]	2020
	Meng et al.[182]	2021
	Hoogeboom et al.[102]	2021
Energy-based Model	Rasul et al. [220]	2021
	Gao et al. [72]	2021
	Yu et al. [316]	2022

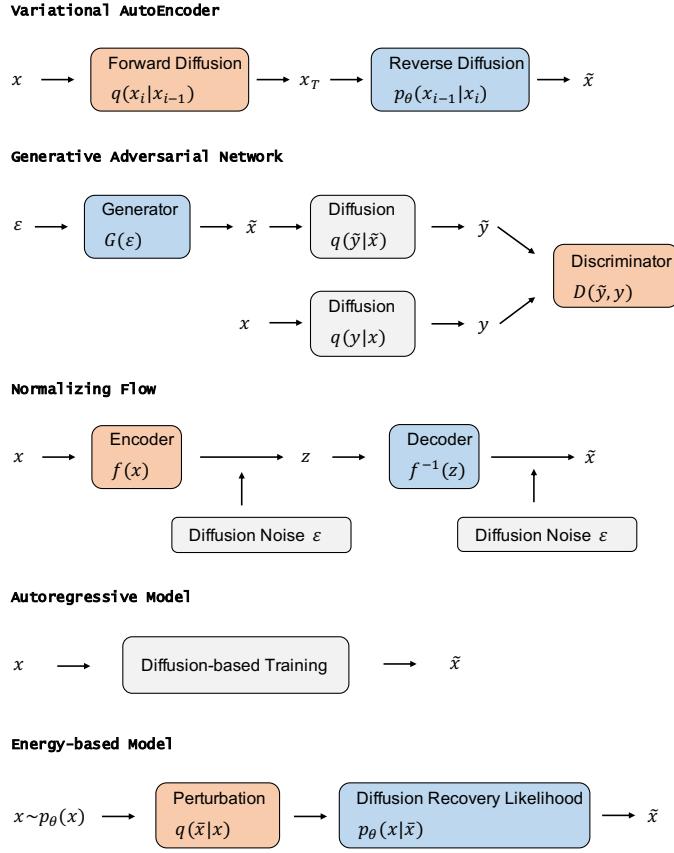


Fig. 3. 关于将扩散模型与其他生成模型结合的作品的插图。

## 6.1 Variational Autoencoders and Connections with Diffusion Models

变分自编码器（Variational Autoencoders）[59, 136, 225]旨在学习一个编码器和一个解码器，将输入数据映射到连续潜变量空间中的值。在这些模型中，嵌入可以被解释为概率生成模型中的潜变量，而概率解码器可以通过参数化的似然函数来定义。另外，假设数据 $\mathbf{x}$ 是通过一些未观测到的潜变量 $\mathbf{z}$ 以条件分布 $p_\theta(\mathbf{x} | \mathbf{z})$ 生成的，而 $q_\phi(\mathbf{z} | \mathbf{x})$ 用来近似推断 $\mathbf{z}$ 。为了保证有效的推断，采用变分贝叶斯方法来最大化证据下界：

$$\mathcal{L}(\phi, \theta; \mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} | \mathbf{x})] \quad (40)$$

以  $\mathcal{L}(\phi, \theta; \mathbf{x}) \leq \log p_\theta(\mathbf{x})$ 。只要参数化的似然函数  $p_\theta(\mathbf{x} | \mathbf{z})$  和参数化的后验逼近  $q_\phi(\mathbf{z} | \mathbf{x})$  在点点之间可计算，并且它们的参数可微分，通过梯度下降法可以最大化ELBO。这种形式允许

灵活选择编码器和解码器模型。通常，这些模型由多层神经网络生成其参数的指数族分布表示。

DDPM可以被概念化为具有固定编码器的分层马尔可夫变分自编码器（VAE）。具体而言，DDPM的前向过程用作编码器，这个过程结构化为线性高斯模型（如Eq. (2)所述）。而DDPM的逆向过程则对应于解码器，该解码器被多个解码步骤共享。解码器内的潜变量与样本数据具有相同大小。

在连续时间设置下，Song et al. (2021) [257]，Huang et al. (2021) [105] 和 Kingma et al. (2021) [133]证明了评分匹配目标可以通过深层分层VAE的证据下界（ELBO）来近似。因此，优化扩散模型可以被视为训练一个无限深的分层VAE——这一发现支持了Score SDE扩散模型可以被解释为分层VAE连续极限的普遍观点。

潜在评分生成模型（LSGM）[268]进一步推进了这一研究方向，通过说明在潜在空间扩散背景下，ELBO可以被视为特殊化的评分匹配目标。虽然ELBO中的交叉熵项是难以处理的，但可以通过将基于评分的生成模型视为无限深的VAE将其转化为可处理的评分匹配目标。

## 6.2 Generative Adversarial Networks and Connections with Diffusion Models

生成对抗网络（GANs）[42, 77, 89] 主要由两个模型组成：一个生成器  $G$  和一个判别器  $D$ 。这两个模型通常由神经网络构建，但也可以以任何将输入数据从一个空间映射到另一个空间的可微分系统进行实现。GANs 的优化可以视为具有值函数  $V(G, D)$  的极小极大优化问题：

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]. \quad (41)$$

生成器  $G$  的目标是生成新的示例并隐性地对数据分布进行建模。判别器  $D$  通常是一个二分类器，用于以可能最大的准确性识别生成的示例和真实示例。优化过程在一个鞍点结束，该鞍点产生生成器的最小值和判别器的最大值。也就是说，GAN优化的目标是实现纳什均衡[223]。在那一点上，可以认为生成器已经捕捉到了真实示例的准确分布。

GAN的一个问题是训练过程中的不稳定性，这主要是由输入数据分布与生成数据分布之间的不重叠引起的。一个解决方案是将噪声注入到判别器输入中，以扩大生成器和判别器分布的支持。利用灵活的扩散模型，Wang等人（2022）[275]在判别器中注入噪声，噪声的调度由扩散模型确定。另一方面，GAN可以提高扩散模型的采样速度。Xiao等人（2021）[290]表明，慢速采样是由于去噪步骤中的高斯假设，这仅适用于小的步长。因此，每个去噪步骤由条件GAN建模，允许更大的步长。

### 6.3 Normalizing Flows and Connections with Diffusion Models

正则化流 [54, 224]是能够生成可计算分布来建模高维数据的生成模型 [56, 134]。正则化流可以将简单的概率分布转化为极其复杂的概率分布，可用于生成模型、强化学习、变分推断等领域。现有的正则化流是基于变量改变公式构建的 [54, 224]。正则化流中的轨迹由微分方程进行建模。在离散时间设置下，正则化流中从数据  $\mathbf{x}$  到潜变量  $\mathbf{z}$  的映射是一系列双射的组合，形式为  $F = F_N \circ F_{N-1} \circ \dots \circ F_1$ 。正则化流中的轨迹  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  满足：

$$\mathbf{x}_i = F_i(\mathbf{x}_{i-1}, \theta), \quad \mathbf{x}_{i-1} = F_i^{-1}(\mathbf{x}_i, \theta) \quad (42)$$

对于所有的  $i \leq N$ 。

类似于连续设定，归一化流通过变量变换公式可以恢复精确的对数似然函数。然而，这种双射要求限制了在实践和理论上对复杂数据建模的可能性[41, 282]。一些工作试图放松这种双射要求[56, 282]。例如，DiffFlow[323]引入了一种将流代模型和扩散模型的优势结合起来的生成建模算法。因此，DiffFlow生成的边界更加锐利，相较于扩散概率模型具有更少的离散化步骤来学习更一般的分布。

隐式非线性扩散模型(INDM)[129]优化了潜在扩散的预编码过程，该过程首先使用归一化流将原始数据编码成潜在空间，并在潜在空间中进行扩散。使用随机微积分的理论，使用归一化流进行数据变换可以看作是学习一个由归一化流确定漂移和扩散系数的非线性SDE。INDM中的得分匹配目标等于扩散模型和归一化流中目标的组合。通过使用非线性扩散过程，INDM可以有效地提高似然函数和采样速度。

### 6.4 Autoregressive Models and Connections with Diffusion Models

自回归模型（ARMs）通过使用概率链规则，将数据的联合分布分解为条件分布的乘积。

$$\log p(\mathbf{x}_{1:T}) = \sum_{t=1}^T \log p(x_t | \mathbf{x}_{<t}) \quad (43)$$

其中  $\mathbf{x}_{<t}$  是  $x_1, x_2, \dots, x_{t-1}$  的简写[12, 142]。最近深度学习的进展促进了各种数据模态的重要进展[27, 181, 237]，例如图像[36, 271]、音频[120, 270]和文本[13, 20, 84, 177, 185]。自回归模型（ARMs）通过使用单个神经网络提供生成能力。从这些模型中采样需要与数据维度相同数量的网络调用。尽管ARMs是有效的密度估计器，但采样是一个持续而耗时的过程，尤其是对于高维数据来说。

另一方面，自回归扩散模型（ARDM）[102]能够生成任意阶数的数据，包括无序自回归模型和离散扩散模型作为特殊情况[6, 103, 248]。ARDM不像ARMs那样在表示上使用因果遮罩，

而是使用与扩散概率模型相似的有效目标进行训练。在测试阶段，ARDM能够并行生成数据，从而适用于各种任意生成任务。

Ment等人（2021）[182]将随机平滑引入自回归生成建模中，以改善样本质量。原始数据分布通过与平滑分布（例如高斯或拉普拉斯核）卷积来进行平滑。平滑后的数据分布通过自回归模型进行学习，然后通过应用基于梯度的去噪方法或引入另一个条件自回归模型来对学习的分布进行去噪处理。通过适当选择平滑程度，所提出的方法可以提高现有自回归模型的样本质量，同时保持合理的似然。

另一方面，自回归条件得分模型（AR-CSM）[184]提出了一种得分匹配方法来建模自回归模型的条件分布。条件分布的得分函数，即 $\nabla_{x_t} \log p(x_t | x_{\leq t})$ ，不需要归一化，因此可以在模型中使用更灵活和先进的神经网络。此外，即使原始数据的维度可能非常高，也可以有效地估计单变量的条件得分函数。在推理中，AR-CSM使用Langevin动力学，只需得分函数即可从密度中采样。

## 6.5 Energy-based Models and Connections with Diffusion Models

能量模型（Energy-based Models，简称EBMs）[28, 50, 61, 67, 70, 71, 79, 82, 83, 132, 141, 144, 187, 192, 211, 226, 291, 328]可以被视为判别器的一种生成型版本[83, 111, 143, 146]，并且可以通过无标签输入数据进行学习。设 $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ 表示一个训练样本， $p_{\theta}(\mathbf{x})$ 表示一个旨在近似 $p_{\text{data}}(\mathbf{x})$ 的概率密度函数。一个能量模型的定义为：

$$p_{\theta}(\mathbf{x}) = \frac{1}{Z_{\theta}} \exp(f_{\theta}(\mathbf{x})), \quad (44)$$

其中 $Z_{\theta} = \int \exp(f_{\theta}(\mathbf{x})) d\mathbf{x}$ 是分区函数，对于高维的 $\mathbf{x}$ 来说，在解析上是难以处理的。对于图像， $f_{\theta}(\mathbf{x})$ 通过一个具有标量输出的卷积神经网络进行参数化。Salimans等人（2021）[234]比较了约束得分模型和基于能量的模型来建模数据分布的得分，并发现当使用可比较的模型结构时，约束得分模型，即基于能量的模型，在性能上可以和无约束模型表现几乎相同。

虽然EBM具有一些理想的性质，但在建模高维数据时仍存在两个挑战。首先，通过最大化似然来学习EBM需要使用MCMC方法从模型中生成样本，这可能计算上非常昂贵。其次，正如[191]所示，使用非收敛的MCMC学习的能量势函数是不稳定的，即来自长期马尔可夫链的样本可能与观测样本显著不同，因此难以评估学到的能量势函数。在最近的研究中，高等人（2021）[72]提出了一种扩散恢复似然方法，以可处理的方式从扩散模型的逆过程中学习EBM的样本。每个EBM都通过恢复似然进行训练，该方法旨在最大化给定更高噪声水平上的嘈杂版本数据的条件概率的情况下，某特定噪声水平上数据的条件概率。EBM最大化恢复似然是因为它比边际似然更容易处理，从条件分布中采样比从边际分布中采样更容易。该模型能够生成高质量的样本，并且长期MCMC样本仍然类似于真实图像。

## 7 APPLICATIONS OF DIFFUSION MODELS

由于其灵活性和强大功能，扩散模型近来被广泛应用于解决各种具有挑战性的现实世界任务。根据任务的不同，我们将这些应用分为六个类别：计算机视觉、自然语言处理、时间数据建模、多模态学习、鲁棒学习和跨学科应用。对于每个类别，我们提供了任务的简要介绍，然后详细解释了如何使用扩散模型来改进性能。表4总结了利用扩散模型进行各种应用的情况。

表 3. 对应于扩散模型的所有应用的概述。

Primary	Secondary	
Computer Vision	Super Resolution, Inpainting, Restoration, Translation, and Editing	[149],[232],[228],[163],[230],[209],[98],[11],[198],[40] [256],[38],[180],[123]
	Semantic Segmentation	[10],[19],[80],[295]
	Video Generation	[94],[100],[310],[322],[245],[96],[283],[210]
	Point Cloud Completion and Generation Anomaly Detection	[334],[167],[172],[159],[320] [289],[73]
Natural Language Generation	Natural Language Generation	[6],[153],[34],[75],[93],[52]
	Time Series Imputation	[262],[1],[201]
	Time Series Forecasting	[221],[1]
Temporal Data Modeling	Waveform Signal Processing	[31],[138]
	Text-to-Image Generation	[7],[216],[231],[189],[87],[65],[229],[126],[269],[9],[299],[321]
	Scene Graph-to-Image Generation	[306]
	Text-to-3D Generation	[296],[155],[207]
Multi-Modal Learning	Text-to-Motion Generation	[263, 322],[130]
	Text-to-Video Generation	[245],[96],[283],[210]
	Text-to-Audio Generation	[208],[301],[286],[148],[260],[106],[131]
	Robust Learning	[190],[312],[17],[274],[285],[259]
Robust Learning	Molecular Graph Modeling	[114],[101],[3],[298],[267],[284],[241],[169]
	Material Design	[293],[170]
	Medical Image Reconstruction	[256],[38],[39],[40],[203],[294]

### 7.1 Unconditional and Conditional Diffusion Models

在我们介绍扩散模型的应用之前，我们先说明一下扩散模型的两个基本应用范式，即无条件扩散模型和条件扩散模型。作为一个生成模型，扩散模型的历史与变分自编码器（VAE）、生成对抗网络（GAN）、流模型等其他生成模型非常相似。它们都首先发展了无条件生成，然后紧随其后发展了条件生成。无条件生成常用于探索生成模型性能的上限，而条件生成更多地涉及应用级内容，因为它可以根据我们的意图来控制生成结果。除了具有良好的生成质量和样本多样性外，扩散模型在可控性方面尤为优越。无条件扩散模型的主要算法已在第2、3、4和5节中进行了充分讨论。接下来，我们将重点讨论条件扩散模型在不同形式条件下的应用，并选择一些典型场景进行演示。

**7.1.1 Conditioning Mechanisms in Diffusion Models.** 利用不同形式的条件来指导扩散模型的生成方向是被广泛使用的，例如标签、分类器、文本、图像、语义地图、图等。然而，其中一些条件是结构复杂的，因此对它们进行条件设定的方法是值得讨论的。主要有四种条件机制，包括串联、基于梯度、交叉注意力和自适应层归一化（adaLN）。串联意味着扩散模型在扩散过程中将信息性指导与中间去噪目标进行串联，例如标签嵌入和语义特征图。基于梯度的机制将任务相关的梯度融入到扩散采样过程中以实现可控的生成。例如，在图像生成中，可以

在噪声图像上训练一个辅助分类器，然后利用梯度将扩散采样过程引导到任意类别标签。交叉注意力在指导和扩散目标之间进行注意力信息传递，通常在去噪网络中以层级方式进行。adaLN机制遵循了GANs中的自适应归一化层的广泛使用[205]，Scalable Diffusion Models[202]探索了将变压器扩散主干网络中的标准层归一化层替换为自适应层归一化。它不直接学习维度上的比例和平移参数，而是通过从时间嵌入和条件的总和中回归它们。

**7.1.2 Condition Diffusion on Labels and Classifiers.** 在生成样本中添加所需属性的直接方法是将扩散过程限制在标签的引导下。然而，当标签有限时，很难使扩散模型充分捕捉整个数据分布。SGGM [308] 提出了一种自我引导的扩散过程，以制作自生成的分层标签集为条件。You等人（2023）[314]证明了大规模扩散模型和半监督学习者通过双重伪训练与少量标签相互受益。Dhariwal和Nichol [51] 提出了“分类器引导”方法，通过使用额外的训练过的分类器提高扩散模型的样本质量。Ho和Salimans [99] 共同训练条件扩散模型和无条件扩散模型，并发现可以结合得到的条件和无条件分数，获得类似于使用分类器引导时样本质量和多样性的权衡。

**7.1.3 Condition Diffusion on Texts, Images, and Semantic Maps.** 最近的研究开始将扩散过程的条件设定在更多语义引导上，例如文本、图像和语义地图，以更好地表达样本中的丰富语义。DiffuSeq [75] 在文本上设定条件，并提出了一种序列到序列的扩散框架，有助于四个自然语言处理任务。SDEdit [180] 对样式化图像进行条件设定，用于图像到图像的翻译，而 LDM [228] 则将这些语义条件与灵活的潜在扩散统一起来。请注意，如果条件和扩散目标属于不同的模态，预对齐[216, 306] 是加强引导扩散的一种实用方法。unCLIP [216] 在文本到图像生成中利用了CLIP潜变量，这些变量已经对齐了图像和文本之间的语义。

**7.1.4 Condition Diffusion on Graphs.** 图结构化数据通常展示出节点之间复杂的关系，因此基于图进行条件生成对扩散模型来说是极具挑战性的。SGDiff [306] 提出了第一个专门针对场景图到图像生成的扩散模型，采用了一种新颖的遮蔽对比预训练方法。这种遮蔽预训练范式具有广泛的适用性，并可扩展到任何粗粒度和细粒度引导的跨模态扩散结构。其他基于图条件的扩散模型主要用于图生成。GeoDiff [298] 以二维分子图作为条件，在保证具有旋转和平移不变性的情况下生成三维分子构型，采用了等变马尔科夫核函数。Luo 等人(2022)[170] 和 DiffSBDD [240] 提出以三维蛋白质图为条件生成三维抗体或分子，采用等变扩散方法。

## 7.2 Computer Vision

**7.2.1 Image Super Resolution, Inpainting, Restoration, Translation, and Editing.** 生成模型已被用于解决各种图像恢复任务，包括超分辨率、修复和翻译[11, 49, 64, 110, 149, 198, 217, 329]。图像超分辨率旨在从低分辨率输入中恢复高分辨率图像，而图像修复则涉及对图像中缺失或损坏的区域进行重建。

有几种方法利用扩散模型来完成这些任务。例如，Super-Resolution via Repeated Refinement (SR3) [232]使用DDPM来实现条件图像生成。SR3通过随机迭代去噪过程进行超分辨率处理。级联扩散模型 (CDM) [98]由多个顺序的扩散模型组成，每个模型生成分辨率更高的图像。SR3和CDM都直接将扩散过程应用于输入图像，这导致了更大的评估步长。为了允许在有限的



Fig. 4. LDM [228] 生成的图像超分辨率结果。

计算资源下训练扩散模型，一些方法[228, 268]使用预训练的自动编码器将扩散过程转移到潜空间中。潜空间扩散模型 (Latent Diffusion Model, LDM)[228]简化了去噪扩散模型的训练和采样过程，而不损失质量。

对于修补任务，RePaint[163]采用了增强的去噪策略，利用重新采样迭代来更好地调节图像条件(见图Fig. 5)。与此同时，Palette[230]采用条件扩散模型为四个图像生成任务（上色，修补，取消裁剪和JPEG恢复）创建了一个统一的框架。图像翻译专注于合成具有特定期望样式

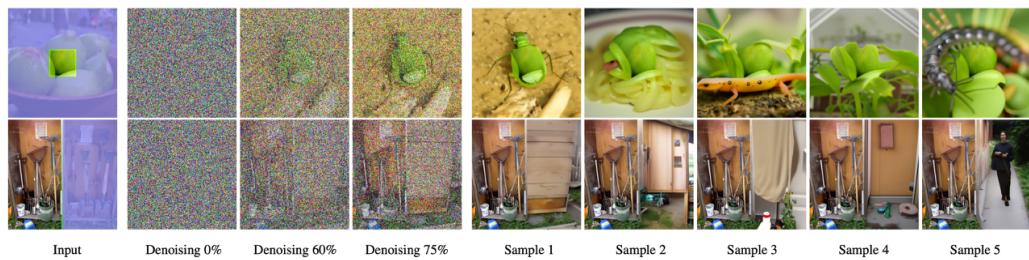


Fig. 5. RePaint [163]生成的图像修复结果。

的图像[110]。SDEdit[180]使用随机微分方程 (SDE) 先验来提高保真度。具体而言，它首先向输入图像添加噪声，然后通过SDE对图像进行去噪处理。去噪扩散恢复模型 (DDRM) [123]利用预训练的去噪扩散生成模型来解决线性逆问题，并展示了DDRM在几个图像数据集上在不同程度的测量噪声下超分辨率，去模糊，修复补全和上色等方面的功能性。有关更多文本到图像扩散模型，请参阅Section 7.4.1。

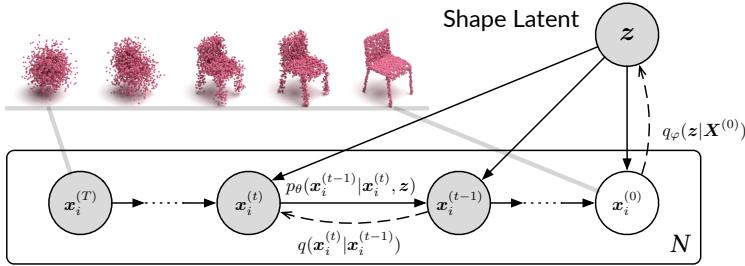


Fig. 6. 点云扩散过程的有向图模型 [167]。

**7.2.2 Semantic Segmentation.** 语义分割旨在根据已建立的物体类别为每个图像像素打标签。生成式预训练可以增强语义分割模型的标签利用率，最近的研究表明通过DDPM学习到的表示具有高级语义信息，对分割任务很有用[10, 80]。利用这些学习到的表示的少样本方法在表现上优于诸如VDVAE[35]和ALAE[206]等替代方法。类似地，Decoder Denoising Pretraining (DDeP)[19]将扩散模型与去噪自编码器[273]结合，对标签高效的语义分割取得了有希望的结果。ODISE[295]利用扩散模型研究开放词汇量的分割任务，并提出了一种新颖的隐式字幕生成器，用于为图像生成字幕，以更好地利用大规模预训练文本到图像扩散模型。

**7.2.3 Video Generation.** 由于视频帧的复杂性和时空连续性，生成高质量的视频仍然是深度学习时代的一个挑战 [305, 317]。最近的研究转向扩散模型以改善生成视频的质量 [100]。例如，灵活扩散模型 (FDM) [94]使用生成模型，使得可以在给定任意其他子集的情况下对任意子集的视频帧进行采样。FDM还包括专门设计用于此目的的架构。此外，残差视频扩散 (RVD) 模型[310]利用自回归的、端到端优化的视频扩散模型。它通过通过逆扩散过程产生的随机残差修正确定性下一帧预测来生成未来帧。更多有关文本到视频扩散模型的信息，请参阅Section 7.4.5。

**7.2.4 Point Cloud Completion and Generation.** 点云是捕捉真实世界对象的关键3D表示形式。然而，由于部分观察或自身遮挡，扫描通常会产生不完整的点云。最近的研究采用扩散模型来解决这一挑战，利用它们来推断缺失的部分，以重建完整的形状。这项工作对于许多下游任务（例如3D重建、增强现实和场景理解）具有重要意义（Lyu等, 2021; Luo等, 2021; Zeng等, 2022）。

Luo等人（2021年）采用了将点云视为热力学系统中的粒子的方法，利用热浴将其从原始分布扩散到噪声分布。同时，Point-Voxel Diffusion (PVD) 模型（Zhou等, 2021）将去噪扩散模型与点-体素表示的3D形状相结合。Point Diffusion-Refinement (PDR) 模型（Lyu等, 2021）使用条件DDPM从部分观察中生成粗糙的补全结果；它还在生成的点云和实际结果之间建立了逐点映射。

**7.2.5 Anomaly Detection.** 异常检测是机器学习[239, 330]和计算机视觉[302]中一个关键且具有挑战性的问题。生成模型已被证明具有强大的机制用于异常检测[73, 92, 289]，通过对正常或健康参考数据进行建模。AnoDDPM[289]利用DDPM对输入图像进行损坏并重建出图像的健康近似。这些方法可能比基于对抗训练的替代方法表现更好，因为它们可以通过有效的采样和稳定的训练方案更好地对较小的数据集进行建模。DDPM-CD[73]通过DDPM将大量的无监督遥感图像纳入训练过程中。使用预训练的DDPM和扩散模型解码器的多尺度表示来检测遥感图像的变化。

### 7.3 Natural Language Generation

自然语言处理旨在理解、建模和管理来自不同来源（如文本或音频）的人类语言。文本生成已成为自然语言处理中最关键和最具挑战性的任务之一[109, 150, 151]。它旨在在人类语言中给定输入数据（例如，序列和关键词）或随机噪声的情况下，生成合理和可读的文本。已经开发了许多基于扩散模型的文本生成方法。离散去噪扩散概率模型（D3PM）[6]引入了类似扩散的生成模型，用于字符级文本生成[30]。通过超越具有均匀转换概率的损坏过程，它推广了多项式扩散模型[103]。大规模自回归语言模型（LMs）能够生成高质量的文本[20, 37, 215, 326]。为了可靠地在实际应用中部署这些LMs，通常希望文本生成过程是可控的。这意味着我们需要生成满足所要求的文本（例如，主题，句法结构）。在文本生成中，控制语言模型的行为而无需重新训练是一个重要的问题[45, 127]。Analog Bits[34]生成模拟比特以表示离散变量，并通过自相条件和非对称时间间隔进一步提高样本质量。

尽管最近的方法已经在控制简单句子属性（例如情感）方面取得了重大成功[139, 303]，但在复杂、细粒度的控制（例如句法结构）方面进展甚微。为了解决更复杂的控制问题，Diffusion-LM[153]提出了一种基于连续扩散的新型语言模型。Diffusion-LM从一系列高斯噪声向量开始，逐步将它们去噪为对应单词的向量。逐步去噪的步骤有助于产生分层连续的潜在表示。这种分层连续的潜变量可以使简单的基于梯度的方法完成复杂的控制。同样地，DiffuSeq[75]还在潜变量空间进行扩散过程，并提出了一种新的条件扩散模型来完成更具挑战性的文本到文本生成任务。Ssd-LM [93] 在自然词汇空间而不是学习的潜变量空间上进行扩散，允许模型融入分类器的指导和模块化控制，而无需调整现成的分类器。CDCD[52]提出用连续的时间和输入空间对分类数据（包括文本）进行扩散模型建模，并设计了一种用于优化的分数插值技术。

### 7.4 Multi-Modal Generation

**7.4.1 Text-to-Image Generation.** 近期，视觉语言模型因其潜在应用广泛受到关注[213]。文本到图像生成是从描述性文本生成相应图像的任务[60, 126, 269]。如Fig. 7所示，给出了一个例子。混合扩散[7]利用了预训练的DDPM[51]和CLIP[213]模型，提出了一种适用于一般用途的基于区域的图像编辑解决方案，该解决方案使用自然语言引导，适用于真实和多样化的图像。

而unCLIP（DALLE-2）[216]则提出了一种两阶段方法，其中包括一个可以生成基于CLIP的图像嵌入并受文本标题条件约束的先验模型，以及一个可以生成受图像嵌入条件约束的图像的扩散解码器。最近，Imagen[231]提出了一种文本到图像的扩散模型，并提供了一个全面的性能评估基准。实验证明，Imagen在与包括VQ-GAN+CLIP[43]、潜在扩散模型[162]和DALL-E 2[216]在内的最先进方法之间表现良好。受到引导扩散模型[51, 99]生成逼真样本和文本到图像模型处理自由形式提示的能力的启发，GLIDE[189]将引导扩散应用于文本条件的图像合成应用中。VQ-Diffusion[87]提出了一种用于文本到图像生成的向量量化扩散模型，该模型消除了单向偏差并避免了累积预测误差。《通用扩散》（Versatile Diffusion）[299]提出了第一个统

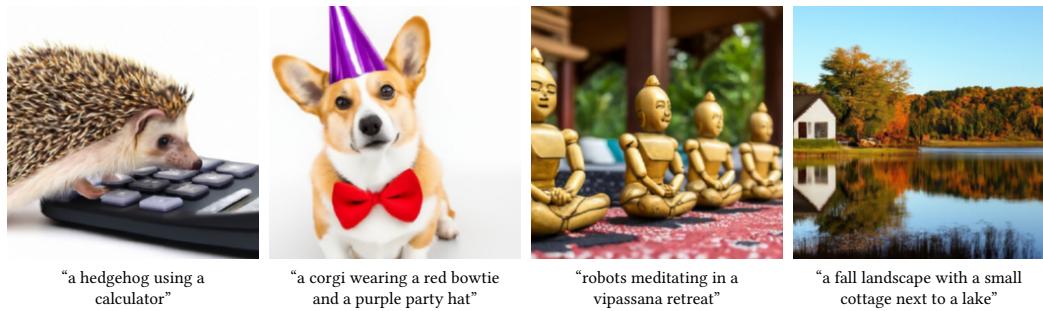


Fig. 7. 由GLIDE [189]生成的文本到图像结果。

一的多流多模态扩散框架，支持图像到文本、图像变化、文本到图像以及文本变化，并可以进一步扩展到其他应用，例如语义风格解缠、图像文本双向生成、潜在图像到文本到图像编辑等等。在《多用途扩散》（Versatile Diffusion）之后，《统一扩散器》（UniDiffuser）[9]提出了基于Transformer的统一扩散模型框架，可以适应多模态数据分布，并同时处理文本到图像、图像到文本以及联合图像-文本生成任务，如Fig. 8所示。在UniDiffuser中，图像和文本首先通过预训练编码器（图像编码器为自编码器，文本编码器为GPT）映射到潜空间。然后，图像和文本的嵌入被串联在一起，并结合指导信息，以控制基于Transformer的扩散生成中不同模态的生成过程。

扩散模型研究的一个新有趣方向是利用预训练的文本到图像扩散模型来实现更复杂或更细粒度的合成结果控制。《梦境摄影棚》（DreamBooth）[229]提出了首个解决主题驱动生成这一新挑战问题的技术，允许用户仅通过拍摄的几张主题图像来重新上下文化主题，修改其属性、原始艺术呈现方式等。该技术通过将预训练的语义先验与自生类特定先验保护损失相结合，学习将唯一标识符与特定主题的输入绑定。

与以文本提示为条件的图像扩散模型不同，《控制网络》（ControlNet）[321]尝试控制预训练的大型扩散模型以支持额外的语义映射，如边缘图、分割图、关键点、形状法线、深度等。如Fig. 9所示，ControlNet提出利用预训练扩散模型的“可训练副本”来避免过拟合。可训练副

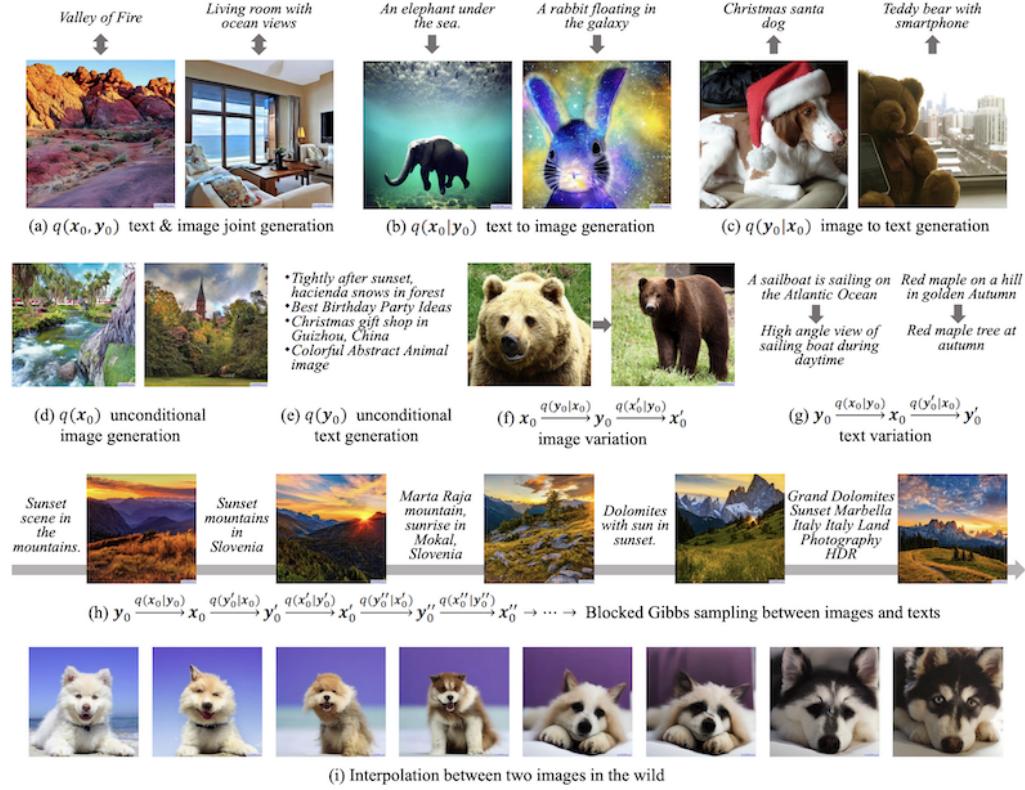


Fig. 8. UniDiffuser [9] 使用基于Transformer的扩散模型处理各种生成任务。

本和原始冻结模型之间通过称为“零卷积”的特殊卷积层连接，其中卷积权重可以通过零初始化进行学习，因为不希望向深层特征添加新的噪声。ControlNet的生成结果示例在Fig. 10中展示。

**7.4.2 Scene Graph-to-Image Generation.** 尽管文本到图像生成模型在自然语言描述方面取得了令人兴奋的进展，但它们在复杂句子中生成许多对象和关系时仍然难以忠实地再现。从场景图生成图像是生成模型的一个重要且具有挑战性的任务[116]。传统方法[95, 116, 154]主要通过预测从场景图到图像布局的方式来生成图像。然而，这种中间表示会丢失场景图中的一些语义信息，最近的扩散模型[228]也无法解决这个问题。SGDiff[306]提出了第一个专门用于从场景图生成图像的扩散模型（见图11），并通过设计的掩码对比预训练方法，在场景图和图像之间实现了全局和局部的语义对齐，学习了一个连续的场景图嵌入来调节潜在的扩散模型。与非扩散和扩散方法相比，SGDiff能够生成更好地表达场景图中密集且复杂关系的图像。然而，高质量的配对场景图-图像数据集稀缺且规模较小，如何利用大规模的文本-图像数据集来增强训练或为更好的初始化提供语义扩散先验仍然是一个开放问题。

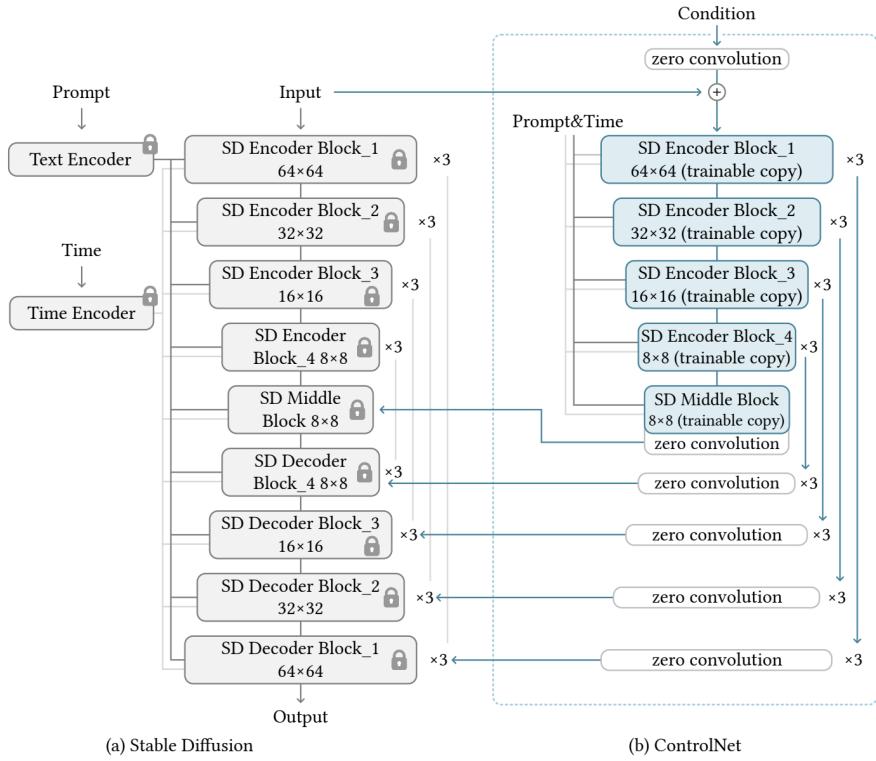


Fig. 9. ControlNet的网络架构[321]包含了一个固定的预训练扩散模型，一个可训练的模型副本以及零卷积。



Fig. 10. ControlNet [321] 用Canny边缘控制稳定扩散。“自动提示”是由基于默认结果图像的模型（BLIP ??）生成的。

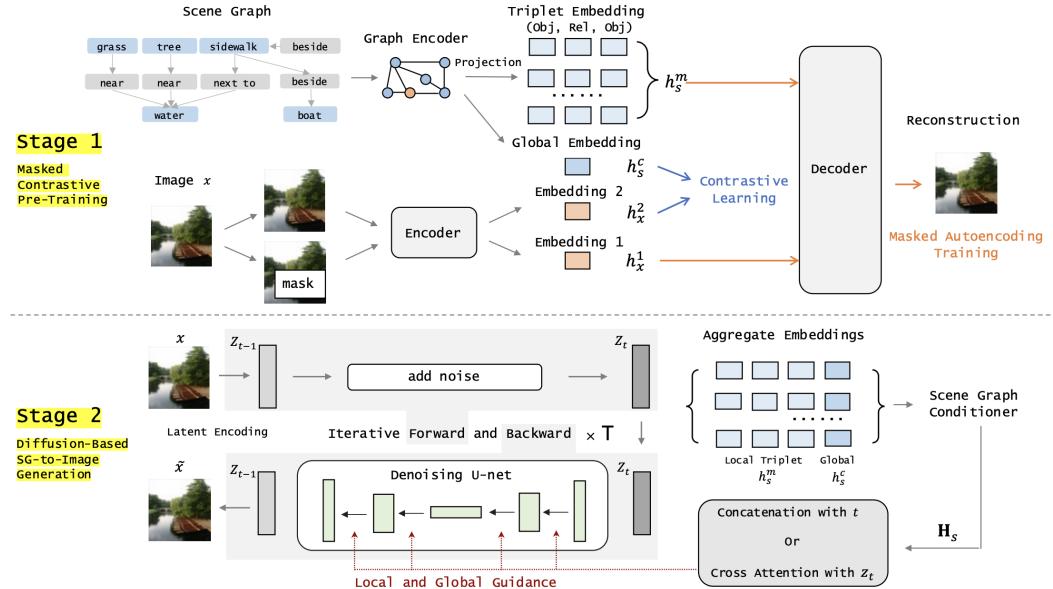


Fig. 11. SGDif [306] 利用遮罩对比性预训练用于基于场景图的图像扩散生成。

**7.4.3 Text-to-3D Generation.** [207] [155] [296] 3D内容生成一直以来受到广泛需求，包括游戏、娱乐和机器人模拟等各种应用领域。通过自然语言来增强3D内容生成，对于初学者和有经验的艺术家都能有很大帮助。DreamFusion [207] 使用一个预训练的2D文本到图像扩散模型来进行文本到3D合成。它通过概率密度扩散损失来优化一个随机初始化的3D模型（神经光辐射场，或者NeRF），该损失利用2D扩散模型作为参数化图像生成器的优化先验。为了获得快速和高分辨率的NeRF优化，Magic3D [155] 提出了一个两阶段扩散框架，建立在级联低分辨率图像扩散先验和高分辨率潜在扩散先验基础上。

**7.4.4 Text-to-Motion Generation.** 人体运动生成是计算机动画中的一项基本任务，应用领域涵盖游戏到机器人技术[322]。生成的运动通常是由关节旋转和位置表示的一系列人体姿势。Motion Diffusion Model (MDM) [263]采用了一种无需分类器的基于扩散的生成模型进行人体运动生成，该模型是基于Transformer的，并结合了运动生成文献中的见解，并通过几何损失对运动的位置和速度进行了规范化。FLAME[130]涉及一种基于Transformer的扩散方法，以更好地处理运动数据，它能够处理可变长度的运动，并能够有效关注自由文本。值得注意的是，它可以在不进行任何微调的情况下编辑运动的部分，包括逐帧和关节级别的编辑。

**7.4.5 Text-to-Video Generation.** 近年来，文本到图像扩散生成领域取得了巨大的进展，这促使了文本到视频生成的发展[96, 245, 283]。Make-A-Video[245]提出通过时空分解扩散模型将基于文本到图像的模型扩展到文本到视频。它利用联合文本图像先验来绕过配对的文本视频数据

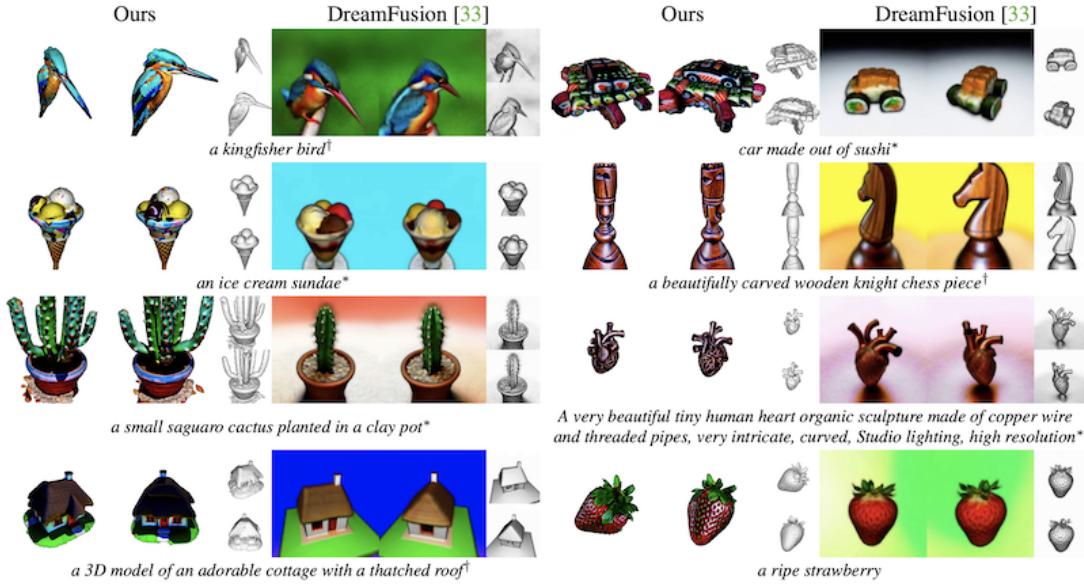


Fig. 12. 比较Magic3D与DreamFusion [155, 207]。

需求，并给出了用于高清、高帧率文本到视频生成的超分辨率策略。Imagen Video[96]通过设计级联视频扩散模型生成高清视频，并将在文本到图像环境中表现良好的一些发现迁移到视频生成领域，包括冻结T5文本编码器和无分类器指导。FateZero[210]是首个使用预训练的文本到图像扩散模型进行时序一致的零样本文本到视频编辑的框架。它融合了DDIM反演和生成过程中的注意力图，以最大限度地保持编辑过程中的动作和结构的一致性。

**7.4.6 Text-to-Audio Generation.** 文本到语音生成是将普通语言文本转化为语音输出的任务[148, 286]。Grad-TTS [208] 提出了一种新颖的文本到语音模型，它采用基于分数的解码器和扩散模型。它通过逐步转化由编码器预测的噪声，并通过单调对齐搜索方法与文本输入进一步对齐。Grad-TTS2 [131] 以一种自适应的方式改进了 Grad-TTS。Diffsound [301] 提供了基于离散扩散模型的非自回归解码器[6, 247]，它在每个单个步骤中预测所有的mel频谱令牌，并在随后的步骤中对预测的令牌进行细化。EdiTTS [260] 利用基于分数的文本到语音模型来改进粗略调整的mel频谱先验。与估计数据密度的梯度不同，ProDiff [106]通过直接预测干净数据来参数化去噪扩散模型。

## 7.5 Temporal Data Modeling

**7.5.1 Time Series Imputation.** 时间序列数据被广泛应用于许多重要的现实应用[63, 196, 305, 327]。然而，时间序列通常由于多种原因包含缺失值，这些原因可能是机械或人为错误引起的[244, 261, 311]。近年来，填补方法在确定性填补[25, 29, 171]和概率性填补[68]方面取得了巨大进展，

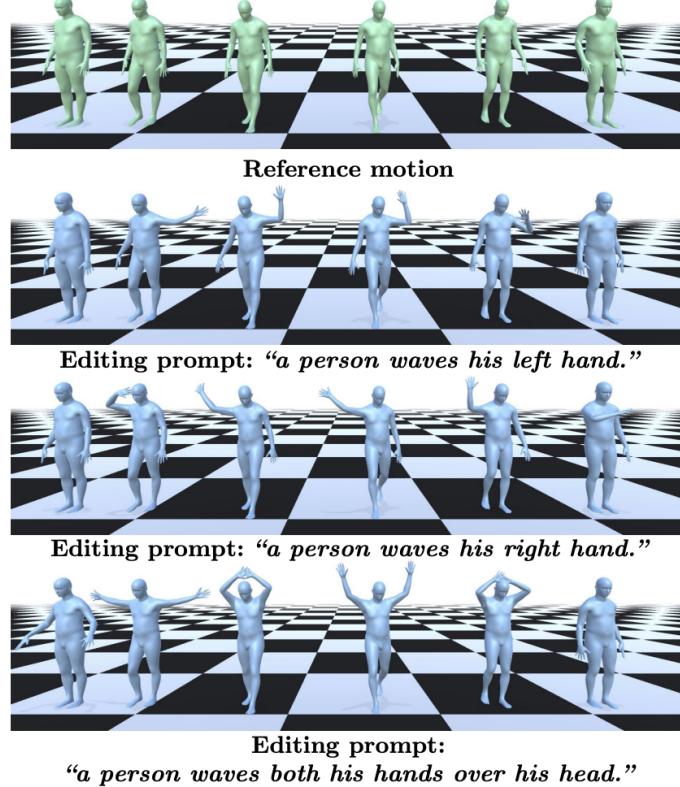


Fig. 13. 基于文本的FLAME动作编辑 [130]。

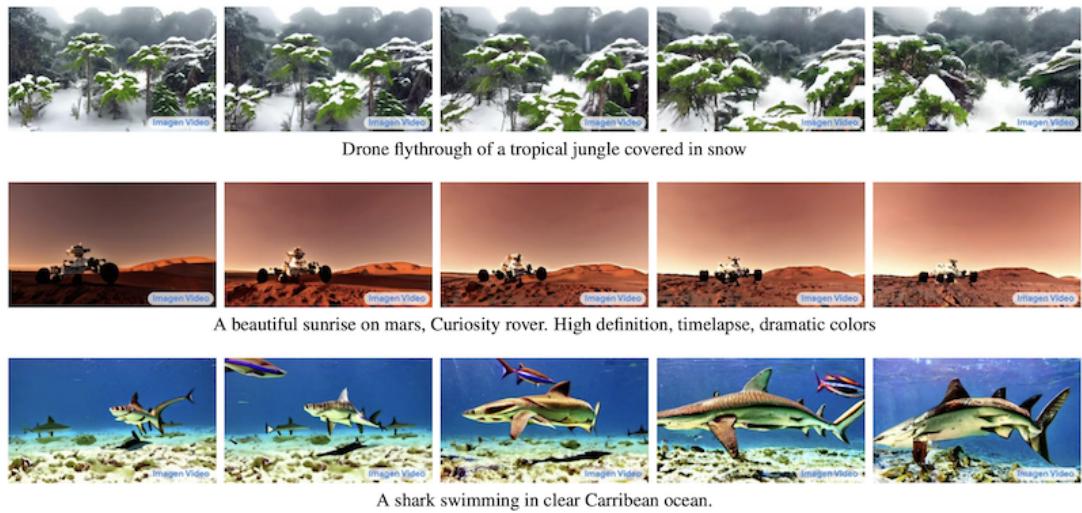


Fig. 14. Imagen Video的文本到视频生成结果 [96]。

包括基于扩散的方法。条件评分扩散模型用于补全（CSDI）[262]提出了一种利用评分扩散模型的新型时间序列补全方法。具体而言，为了利用时间数据内部的相关性，它采用自监督训练的形式来优化扩散模型。在一些真实世界数据集的应用中，它表现出对先前方法的优越性。控制随机微分方程（CSDE）[201]提出了一种用于建模具有神经控制的随机动力学的新型概率性框架。结构化状态空间扩散（SSSD）[1]将条件扩散模型和结构化状态空间模型[86]集成在一起，特别捕捉时间序列中的长期依赖关系。它在时间序列补全和预测任务中表现良好。

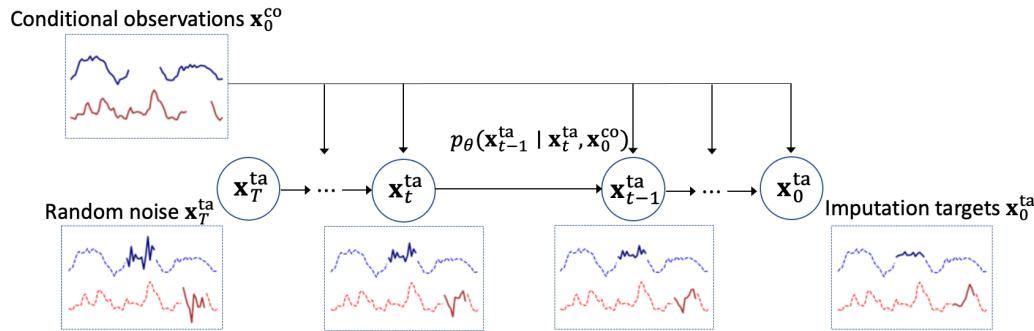


Fig. 15. CSDI 时间序列插值的步骤 [262].

**7.5.2 Time Series Forecasting.** 时间序列预测是在一段时间内预测或预测未来价值的任务。神经方法最近已广泛应用于解决单变量点预测方法[195]或单变量概率方法[235]的预测问题。在多变量设置中，我们也有点预测方法[152]以及概率方法，它们明确地使用高斯联合分布模型数据分布[236]，GANs[313]或正则化流[222]来建模。TimeGrad [221] 提出了一种用于预测多元概率时间序列的自回归模型，它通过估计其梯度在每个时间步骤中从数据分布中抽样。它利用扩散概率模型，这些模型与得分匹配和基于能量的方法密切相关。具体来说，它通过优化数据似然的变分界限学习梯度，并在推理时使用 Langevin 抽样[252]将白噪声转化为感兴趣的分布样本的马尔可夫链。

**7.5.3 Waveform Signal Processing.** 在电子学、声学和一些相关领域中，信号的波形被表示为其随时间的图形形状，与其时间和幅度尺度无关。

WaveGrad [31] 引入了一个条件模型用于波形生成，它估计数据密度的梯度。它接收高斯白噪声信号作为输入，并通过基于梯度的采样器迭代地改进信号。

WaveGrad 自然地通过调整改进步骤的数量在推断速度和样本质量之间进行平衡，并在音频质量方面使非自回归模型和自回归模型之间建立了联系。

DiffWave [138] 提出了一种多功能且有效的扩散概率模型，用于条件或无条件波形生成。该模型是非自回归的，并通过优化数据似然的变体来进行高效训练。

此外，它在不同的波形生成任务中产生高保真度音频，例如类条件生成和无条件生成。

## 7.6 Robust Learning

Robust learning是一种防御方法的类别，它有助于使学习网络对抗性扰动或噪声具有鲁棒性 [17, 190, 206, 274, 285, 312]。虽然对抗训练 [174] 被视为抵御对图像分类器的对抗攻击的标准防御方法，但对抗净化作为另一种防御方法表现出显著的性能 [312]，它通过独立的净化模型将遭受攻击的图像净化成干净的图像。给定一个对抗性示例，DiffPure [190] 使用前向扩散过程将其扩散一小部分噪声，并通过逆向生成过程恢复干净图像。自适应去噪净化（ADP）[312] 表明，通过使用去噪评分匹配 [272] 训练的EBM可以在几个步骤内有效地净化受攻击的图像。它进一步提出了一种有效的随机净化方案，在净化之前向图像注入随机噪声。Projected Gradient Descent（PGD）[17] 提出了一种新颖的基于随机扩散的前处理鲁棒化方法，旨在成为一种与模型无关的对抗性防御，并产生高质量的去噪结果。此外，一些工作提出了应用导向扩散过程进行高级对抗净化 [274, 285]。

## 7.7 Interdisciplinary Applications

**7.7.1 Drug Design and Life Science.** 图神经网络[90, 288, 307, 333]及其相应的表示学习[91]技术在许多领域取得了巨大成功[15, 264, 287, 297, 304, 336]，包括在模拟分子/蛋白质的各种任务中，从性质预测[62, 74]到分子/蛋白质生成[112, 119, 169, 242]，其中分子自然地由节点-边图表示。

一方面，近期的研究提出了专门为具有生物医学或物理学见解的分子/蛋白质预训练GNN/transformer[166, 332]，并取得了显著的结果[157, 319]。另一方面，更多的研究开始利用基于图的扩散模型来增强分子或蛋白质的生成。Torsional diffusion[114]提出了一种新的扩散框架，利用扩散过程在超空间中进行torsion角度的操作，并结合外在-内在评分模型。

GeoDiff[298]证明了与等变Markov核一起演化的马尔可夫链可以产生一个不变分布，并进一步设计了用于保持所需等变性质的Markov核的块。还有其他研究将等变性质应用于3D分子生成[101]和蛋白质生成[3, 14]中。

受经典力场方法模拟分子动力学的启发，ConfGF[241]直接估计分子构象生成中原子坐标的对数密度的梯度场。

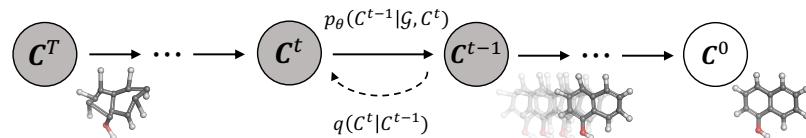


Fig. 16. GeoDiff中的分子到构象扩散过程 [298]。

最近，针对目标蛋白质，通过扩散模型来设计能够紧密结合目标的三维小分子药物开始受到推广。TargetDiff使用目标蛋白质作为引导信息，通过明确模拟蛋白质和分子在三维空间中的相互作用，逐步生成分子。此外，该研究还发现，经过训练的扩散生成模型可以作为评  
Manuscript submitted to ACM

分函数，提高结合亲和力的预测准确性。在一些目标蛋白质上，该模型能够与以前的自回归生成模型（如Pocket2Mol）相竞争，表现出卓越的性能。

也有一些研究使用扩散模型进行蛋白质生成，例如DiffAb。DiffAb首次提出了一种基于扩散的三维抗体设计框架，可以对决定抗体互补性的互补决定区（CDRs）的序列和结构进行建模。其多分支扩散模型框架如图17所示。实验证明，DiffAb可用于各种抗体设计任务，例如联合生成序列-结构，设计具有固定框架的CDRs以及优化抗体。SMCDiff [267]首先通过E(3)-等变

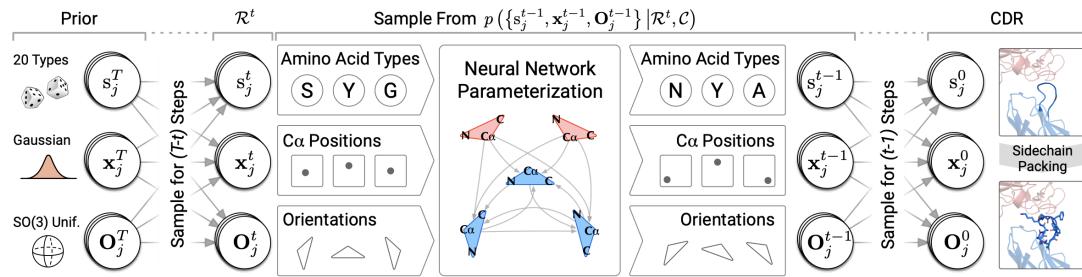


Fig. 17. DiffAb的扩散框架 [170].

图神经网络学习多样且更长的蛋白质主链结构的分布，然后在给定基序的情况下，从该分布中高效地采样支架。生成结果表明，设计的主链与AlphaFold2预测的结构良好对齐。

**7.7.2 Material Design.** 固态材料是许多关键技术的基础，Butler等人(2018) [21]。晶体扩散变分自编码器(CDVAE) [293]通过提出一个噪声条件评分网络，将稳定性作为归纳偏差加入，同时利用置换、平移、旋转和周期性不变性属性。Luo等人(2022) [170]通过等变扩散对互补决定区的序列和结构进行建模，并明确地针对特定抗原结构以原子分辨率生成抗体。

**7.7.3 Medical Image Reconstruction.** 逆问题是从观测测量中恢复未知信号的问题，在计算机断层扫描（CT）和磁共振成像（MRI）的医学图像重建中是一个重要问题[1-5]。Song等人（2021）[1]利用基于评分的生成模型重建与先验知识和观测测量一致的图像。Chung等人（2022）[2]通过使用去噪评分匹配训练连续的时间相关评分函数，在评估阶段通过数值SDE求解器和数据一致性步骤之间迭代，进行重建。Peng等人（2022）[3]根据观察到的k空间信号逐渐引导逆扩散过程进行MR重建，并提出了一种粗到精的采样算法以提高采样效率。

## 8 FUTURE DIRECTIONS

扩散模型的研究仍处于初级阶段，从理论和实证角度都有很大的改进潜力。正如前面几节讨论的，关键的研究方向包括高效采样和改进可能性，以及探索扩散模型如何处理特殊的数据结构，如何与其他类型的生成模型进行接口，并量身定制适用于各种应用的模型。此外，我们预见到未来的扩散模型研究可能会涵盖以下几个方面。

**重新审视假设.** 扩散模型中的许多典型假设需要重新审视和分析。例如，假设扩散模型的前向过程完全抹去了数据中的任何信息，使其等同于先验分布的假设并不总是成立。事实上，在有限时间内完全删除信息是不可实现的。了解何时停止前向扰动过程以在采样效率和样本质量之间取得平衡非常有趣[69]。Schrödinger桥和最优输运的最新进展[33, 46, 48, 243, 250]提供了有希望的替代解决方案，为扩散模型提出了新的公式，能够在有限时间内收敛到指定的先验分布。

**理论理解.** 扩散模型作为一种强大的框架已经出现，尤其是作为唯一一种无需采用对抗训练就能在大多数应用中与生成对抗网络（GANs）相媲美的框架。充分发挥这种潜力的关键是理解为什么和何时扩散模型比其他类型的生成模型更适用于特定任务。重要的是要确定哪些基本特征区分了扩散模型与其他类型的生成模型（如变分自编码器、基于能量的模型或自回归模型）。理解这些差异将有助于阐明为什么扩散模型能够生成出优质样本并取得最佳可能性。同样重要的是需要发展理论指导，以系统地选择和确定扩散模型的各种超参数。

**潜在表示** 与变分自编码器或生成对抗网络不同，扩散模型对数据在其潜在空间中的表征能力较弱。因此，它们不能轻松地用于基于语义表示的数据操作等任务。此外，由于扩散模型中的潜在空间通常具有与数据空间相同的维度，采样效率受到负面影响，模型可能无法很好地学习表示方案[113]。

**AIGC和扩散基础模型.** 从稳定扩散到ChatGPT，人工智能生成内容（AIGC）在学术界和工业界引起了很大关注。生成预训练是GPT-1/2/3/4[194, 197, 214, 215]和（视觉）ChatGPT[281]的核心技术，展现了有希望的生成性能和令人惊讶的新能力[279]，配备了大型语言模型[266]和视觉基础模型[18, 315, 318]。将生成预训练（仅解码器）从GPT系列转移到扩散模型类，评估在大规模下基于扩散的生成性能，并分析扩散基础模型的新能力是非常有趣的研究方向。此外，扩散基础模型可能引发更多迷人的AIGC应用。

## 9 CONCLUSION

我们全面地介绍了扩散模型的各个方面。我们首先对三种基本的表述进行了自包含的介绍：DDPMs、SGMs和Score SDEs。然后，我们讨论了近期提高扩散模型的努力，重点介绍了三个主要的方向：采样效率、似然最大化以及适用于具有特殊结构数据的新技术。我们还探讨了扩散模型与其他生成模型之间的联系，并概述了将两者结合的潜在好处。通过对六个领域的应用调查，我们展示了扩散模型具有广泛潜力。最后，我们概述了未来研究的可能方向。

## REFERENCES

- [1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. 2022. Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. *arXiv preprint arXiv:2208.09399* (2022).

- [2] Tomer Amit, Eliya Nachmani, Tal Shahrabany, and Lior Wolf. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390* (2021).
- [3] Namrata Anand and Tudor Achim. 2022. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2205.15019* (2022).
- [4] Brian DO Anderson. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications* 12, 3 (1982), 313–326.
- [5] Uri M Ascher and Linda R Petzold. 1998. *Computer methods for ordinary differential equations and differential-algebraic equations*. Vol. 61. Siam.
- [6] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*.
- [7] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*. 18208–18218.
- [8] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. 2021. Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models. In *International Conference on Learning Representations*.
- [9] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. 2023. One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale. *arXiv preprint arXiv:2303.06555* (2023).
- [10] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Label-Efficient Semantic Segmentation with Diffusion Models. In *International Conference on Learning Representations*.
- [11] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. 2021. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606* (2021).
- [12] Samy Bengio and Yoshua Bengio. 2000. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Trans. Neural Networks Learn. Syst.* (2000).
- [13] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research* 3 (2003), 1137–1155.
- [14] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. 2000. The protein data bank. *Nucleic acids research* 28, 1 (2000), 235–242.
- [15] Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V Chawla. 2021. Graph Barlow Twins: A self-supervised representation learning framework for graphs. *arXiv preprint arXiv:2106.02466* (2021).
- [16] Mikolaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.
- [17] Tsachi Blau, Roy Ganz, Bahjat Kawar, Alex Bronstein, and Michael Elad. 2022. Threat Model-Agnostic Adversarial Defense using Diffusion Models. *arXiv preprint arXiv:2207.08089* (2022).
- [18] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
- [19] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. 2022. Denoising Pretraining for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4175–4186.
- [20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.
- [21] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature* 559, 7715 (2018), 547–555.
- [22] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. 2020. Learning gradient fields for shape generation. In *European Conference on Computer Vision*. Springer, 364–381.
- [23] Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. 2022. A Continuous Time Framework for Discrete Denoising Models. *arXiv preprint arXiv:2205.14987* (2022).
- [24] Chentao Cao, Zhuo-Xu Cui, Shaonan Liu, Dong Liang, and Yanjie Zhu. 2022. High-Frequency Space Diffusion Models for Accelerated MRI. *arXiv preprint arXiv:2208.05481* (2022).

- [25] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, Vol. 31.
- [26] Nicholas Carlini, Florian Tramer, Krishnamurthy Dvijotham1, and Kolter J. Zico. 2022. (Certified!!) Adversarial Robustness for Free! *arXiv preprint arXiv:2206.10550* (2022).
- [27] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11315–11325.
- [28] Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. 2020. Your GAN is Secretly an Energy-based Model and You Should use Discriminator Driven Latent Sampling. *arXiv preprint arXiv:2003.06060* (2020).
- [29] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 1–12.
- [30] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Philipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005* (2013).
- [31] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. WaveGrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713* (2020).
- [32] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2018. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366* (2018).
- [33] Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. 2021. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations*.
- [34] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. 2022. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. *arXiv preprint arXiv:2208.04202* (2022).
- [35] Rewon Child. 2020. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. In *International Conference on Learning Representations*.
- [36] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. *CoRR* abs/1904.10509 (2019). arXiv:1904.10509 <http://arxiv.org/abs/1904.10509>
- [37] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling Language Modeling with Pathways. (2022).
- [38] Hyungjin Chung, Eun Sun Lee, and Jong Chul Ye. 2022. MR Image Denoising and Super-Resolution Using Regularized Reverse Diffusion. *arXiv preprint arXiv:2203.12621* (2022).
- [39] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *IEEE Conference on Computer Vision and Pattern Recognition*. 12413–12422.
- [40] Hyungjin Chung and Jong Chul Ye. 2022. Score-based diffusion models for accelerated MRI. *Medical Image Analysis* (2022), 102479.
- [41] Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. 2020. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International Conference on Machine Learning*. 2133–2143.
- [42] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine* 35, 1 (2018), 53–65.
- [43] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583* (2022).
- [44] Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Gokberk Elmas, Muzaffer Özbeý, Alper Güngör, and Tolga Çukur. 2022. Adaptive Diffusion Priors for Accelerated MRI Reconstruction. *arXiv preprint arXiv:2207.05876* (2022).
- [45] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.
- [46] Valentin De Bortoli, Arnaud Doucet, Jeremy Heng, and James Thornton. 2021. Simulating diffusion bridges with score matching. *arXiv preprint arXiv:2111.07243* (2021).
- [47] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. 2022. Riemannian score-based generative modeling. *arXiv preprint arXiv:2202.02763* (2022).
- [48] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. 2021. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, Vol. 34. 17695–17709.

- [49] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
- [50] Guillaume Desjardins, Yoshua Bengio, and Aaron C Courville. 2011. On tracking the partition function. In *Advances in Neural Information Processing Systems*. 2501–2509.
- [51] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, Vol. 34. 8780–8794.
- [52] Sander Dieleman, Laurent Sartran, Arman Roshnai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. 2022. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089* (2022).
- [53] Laurent Dinh, David Krueger, and Yoshua Bengio. 2015. Nice: Non-linear independent components estimation. *ICLR 2015 Workshop Track* (2015).
- [54] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803* (2016).
- [55] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using Real NVP. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HlkpbnH9lx>
- [56] Laurent Dinh, Jascha Sohl-Dickstein, Hugo Larochelle, and Razvan Pascanu. 2019. A RAD approach to deep mixture models. *arXiv preprint arXiv:1903.07714* (2019).
- [57] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. 2021. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In *International Conference on Learning Representations*.
- [58] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. 2022. GENIE: Higher-Order Denoising Diffusion Solvers. *Advances in Neural Information Processing Systems* (2022).
- [59] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
- [60] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936* (2022).
- [61] Yilun Du and Igor Mordatch. 2019. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689* (2019).
- [62] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, Vol. 28.
- [63] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. *arXiv preprint arXiv:2106.14112* (2021).
- [64] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 12873–12883.
- [65] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. 2022. Frido: Feature Pyramid Diffusion for Complex Scene Image Synthesis. *arXiv preprint arXiv:2208.13753* (2022).
- [66] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society* 29, 4 (2016), 983–1049.
- [67] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. 2016. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852* (2016).
- [68] Vincent Fortuin, Dmitry Baranchuk, Gunnar Ratsch, and Stephan Mandt. 2020. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*. PMLR, 1651–1661.
- [69] Giulio Franzese, Simone Rossi, Lixuan Yang, Alessandro Finomore, Dario Rossi, Maurizio Filippone, and Pietro Michiardi. 2022. How much is enough? a study on diffusion times in score-based generative models. *arXiv preprint arXiv:2206.05173* (2022).
- [70] Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. 2018. Learning generative convnets via multi-grid modeling and sampling. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9155–9164.
- [71] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. 2020. Flow contrastive estimation of energy-based models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7518–7528.
- [72] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. 2020. Learning energy-based models by diffusion recovery likelihood. *arXiv preprint arXiv:2012.08125* (2020).

- [73] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. 2022. Remote Sensing Change Detection (Segmentation) using Denoising Diffusion Probabilistic Models. *arXiv e-prints* (2022), arXiv–2206.
- [74] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*. 1263–1272.
- [75] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations*.
- [76] Wenbo Gong and Yingzhen Li. 2021. Interpreting diffusion score matching using normalizing flow. *arXiv preprint arXiv:2107.10072* (2021).
- [77] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Vol. 27. 139–144.
- [78] Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks*, Vol. 2. 729–734.
- [79] Anirudh Goyal Alias Parth Goyal, Nan Rosemary Ke, Surya Ganguli, and Yoshua Bengio. 2017. Variational walkback: Learning a transition operator as a stochastic recurrent net. In *Advances in Neural Information Processing Systems*. 4392–4402.
- [80] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. 2022. Diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*.
- [81] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. 2019. Scalable Reversible Generative Models with Free-form Continuous Dynamics. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rJxgknCcK7>
- [82] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2019. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. *arXiv preprint arXiv:1912.03263* (2019).
- [83] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. 2020. Cutting out the Middle-Man: Training and Evaluating Energy-Based Models without Sampling. *arXiv preprint arXiv:2002.05616* (2020).
- [84] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *CoRR abs/1308.0850* (2013). arXiv:1308.0850 <http://arxiv.org/abs/1308.0850>
- [85] Ulf Grenander and Michael I Miller. 1994. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 56, 4 (1994), 549–581.
- [86] Albert Gu, Karan Goel, and Christopher Re. 2021. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*.
- [87] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10696–10706.
- [88] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 2023. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In *International Conference on Learning Representations*.
- [89] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. 2021. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering* (2021).
- [90] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 1025–1035.
- [91] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584* (2017).
- [92] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. 2022. ADBench: Anomaly Detection Benchmark. *arXiv preprint arXiv:2206.09426* (2022).
- [93] Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2022. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. *arXiv preprint arXiv:2210.17432* (2022).
- [94] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. 2022. Flexible Diffusion Modeling of Long Videos. *arXiv preprint arXiv:2205.11495* (2022).
- [95] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. 2020. Learning canonical representations for scene graph to image generation. In *European Conference on Computer Vision*. 210–227.

- [96] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).
- [97] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, Vol. 33. 6840–6851.
- [98] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.* 23 (2022), 47–1.
- [99] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
- [100] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *arXiv preprint arXiv:2204.03458* (2022).
- [101] Emiel Hoogeboom, Victor Garcia Satorras, Clement Vignac, and Max Welling. 2022. Equivariant Diffusion for Molecule Generation in 3D. *arXiv e-prints* (2022), arXiv–2203.
- [102] Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. 2021. Autoregressive Diffusion Models. In *International Conference on Learning Representations*.
- [103] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, Vol. 34. 12454–12465.
- [104] Chin-Wei Huang, Milad Aghajohari, A. Bose, P. Panangaden, and Aaron C. Courville. 2022. Riemannian Diffusion Models.
- [105] Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. 2021. A variational perspective on diffusion-based generative models and score matching. In *Advances in Neural Information Processing Systems*, Vol. 34. 22863–22876.
- [106] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. ProDiff: Progressive Fast Diffusion Model For High-Quality Text-to-Speech. *arXiv preprint arXiv:2207.06389* (2022).
- [107] Michael F Hutchinson. 1989. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation* 18, 3 (1989), 1059–1076.
- [108] Aapo Hyvärinen. 2005. Estimation of Non-Normalized Statistical Models by Score Matching. *J. Mach. Learn. Res.* 6 (2005), 695–709.
- [109] Touseef Iqbal and Shaima Qureshi. 2020. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences* (2020).
- [110] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.
- [111] Long Jin, Justin Lazarow, and Zhuowen Tu. 2017. Introspective classification with convolutional nets. In *Advances in Neural Information Processing Systems*, Vol. 30. 823–833.
- [112] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*. 2323–2332.
- [113] Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. 2022. Subspace diffusion generative models. *arXiv preprint arXiv:2205.01490* (2022).
- [114] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. 2022. Torsional Diffusion for Molecular Conformer Generation. *arXiv preprint arXiv:2206.01729* (2022).
- [115] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. 2022. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*. PMLR, 10362–10383.
- [116] Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1219–1228.
- [117] Alexia Jolicœur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. 2021. Gotta Go Fast When Generating Data with Score-Based Models. (2021).
- [118] Alexia Jolicœur-Martineau, Remi Piche-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. 2021. Adversarial score matching and improved sampling for image generation. *ArXiv abs/2009.05475* (2021).
- [119] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

- [120] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient Neural Audio Synthesis. In *International Conference on Machine Learning*. 2410–2419.
- [121] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. *arXiv preprint arXiv:2206.00364* (2022).
- [122] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [123] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793* (2022).
- [124] Bahjat Kawar, Roy Ganz, and Michael Elad. 2022. Enhancing diffusion-based image synthesis with robust classifier guidance. *arXiv preprint arXiv:2208.08664* (2022).
- [125] Bahjat Kawar, Gregory Vaksman, and Michael Elad. 2021. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1866–1875.
- [126] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-Based Real Image Editing with Diffusion Models. *arXiv preprint arXiv:2210.09276* (2022).
- [127] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).
- [128] Boah Kim, Inhwu Han, and Jong Chul Ye. 2021. Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. *arXiv preprint arXiv:2112.05149* (2021).
- [129] Dongjun Kim, Byeonghu Na, Se Jung Kwon, Dongsoo Lee, Wanmo Kang, and Il-chul Moon. 2022. Maximum Likelihood Training of Implicit Nonlinear Diffusion Model. In *Advances in Neural Information Processing Systems*.
- [130] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2022. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349* (2022).
- [131] Sungwon Kim, Heeseung Kim, and Sungroh Yoon. 2022. Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data. *arXiv preprint arXiv:2205.15370* (2022).
- [132] Taesup Kim and Yoshua Bengio. 2016. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439* (2016).
- [133] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational diffusion models. In *Advances in Neural Information Processing Systems*, Vol. 34. 21696–21707.
- [134] Diederik P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039* (2018).
- [135] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [136] Diederik P Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392.
- [137] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.
- [138] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).
- [139] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367* (2020).
- [140] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).
- [141] Rithesh Kumar, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. 2019. Maximum Entropy Generators for Energy-Based Models. *arXiv preprint arXiv:1901.08508* (2019).
- [142] Hugo Larochelle and Iain Murray. 2011. The Neural Autoregressive Distribution Estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS*.
- [143] Justin Lazarow, Long Jin, and Zhuowen Tu. 2017. Introspective neural networks for generative modeling. In *Proceedings of the IEEE International Conference on Computer Vision*. 2774–2783.

- [144] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc’ Aurelio Ranzato, and Fujie Huang. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).
- [145] Jin Sub Lee and Philip M Kim. 2022. ProteinSGM: Score-based generative modeling for de novo protein design. *bioRxiv* (2022).
- [146] Kwonjoon Lee, Weijian Xu, Fan Fan, and Zhuowen Tu. 2018. Wasserstein introspective neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3702–3711.
- [147] Seul Lee, Jaehyeong Jo, and Sung Ju Hwang. 2022. Exploring Chemical Space with Score-based Out-of-distribution Generation. *arXiv preprint arXiv:2206.07632* (2022).
- [148] Alon Levkovich, Eliya Nachmani, and Lior Wolf. 2022. Zero-Shot Voice Conditioning for Denoising Diffusion TTS Models. *arXiv preprint arXiv:2206.02246* (2022).
- [149] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhi hai Xu, Qi Li, and Yue ting Chen. 2022. SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models. *Neurocomputing* 479 (2022), 47–59.
- [150] Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Textbox: A unified, modularized, and extensible framework for text generation. *arXiv preprint arXiv:2101.02046* (2021).
- [151] Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311* (2021).
- [152] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhua Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*, Vol. 32.
- [153] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. *arXiv preprint arXiv:2205.14217* (2022).
- [154] Yikang Li, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, and Xiaogang Wang. 2019. Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems* 32 (2019).
- [155] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2022. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv preprint arXiv:2211.10440* (2022).
- [156] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2021. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *International Conference on Learning Representations*.
- [157] Shengchao Liu, Hongyu Guo, and Jian Tang. 2023. Molecular geometry pretraining with se (3)-invariant denoising distance matching. In *International Conference on Learning Representations*.
- [158] Xingchao Liu, Lemeng Wu, Mao Ye, et al. 2023. Learning Diffusion Bridges on Constrained Domains. In *International Conference on Learning Representations*.
- [159] Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. 2022. Let us Build Bridges: Understanding and Extending Diffusion Generative Models. *arXiv preprint arXiv:2208.14699* (2022).
- [160] Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M De Sa. 2020. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems* 33 (2020), 17548–17558.
- [161] Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Maximum Likelihood Training for Score-based Diffusion ODEs by High Order Denoising Score Matching. In *International Conference on Machine Learning*. 14429–14460.
- [162] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *arXiv preprint arXiv:2206.00927* (2022).
- [163] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11461–11471.
- [164] Eric Luhman and Troy Luhman. 2021. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388* (2021).
- [165] Calvin Luo. 2022. Understanding Diffusion Models: A Unified Perspective. *arXiv preprint arXiv:2208.11970* (2022).
- [166] Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. 2023. One transformer can understand both 2d & 3d molecular data. In *International Conference on Learning Representations*.
- [167] Shitong Luo and Wei Hu. 2021. Diffusion probabilistic models for 3d point cloud generation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2837–2845.

- [168] Shitong Luo and Wei Hu. 2021. Score-based point cloud denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4583–4592.
- [169] Shitong Luo, Chence Shi, Minkai Xu, and Jian Tang. 2021. Predicting molecular conformation via dynamic graph score matching. In *Advances in Neural Information Processing Systems*, Vol. 34. 19784–19795.
- [170] Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. 2022. Antigen-specific antibody design and optimization with diffusion-based generative models. *bioRxiv* (2022).
- [171] Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. 2018. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, Vol. 31.
- [172] Zhaoyang Lyu, Zhifeng Kong, XU Xudong, Liang Pan, and Dahua Lin. 2021. A Conditional Point Diffusion-Refinement Paradigm for 3D Point Cloud Completion. In *International Conference on Learning Representations*.
- [173] Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. 2022. Accelerating Diffusion Models via Early Stop of the Diffusion Process. *arXiv preprint arXiv:2205.12524* (2022).
- [174] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.
- [175] Emile Mathieu and Maximilian Nickel. 2020. Riemannian continuous normalizing flows. *Advances in Neural Information Processing Systems* 33 (2020), 2503–2515.
- [176] Siyuan Mei, Fuxin Fan, and Andreas Maier. 2022. Metal Inpainting in CBCT Projections Using Score-based Generative Model. *arXiv preprint arXiv:2209.09733* (2022).
- [177] Gábor Melis, Chris Dyer, and Phil Blunsom. 2018. On the State of the Art of Evaluation in Neural Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ByJHuTgA>
- [178] Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. 2022. Concrete Score Matching: Generalized Score Matching for Discrete Data. In *Advances in Neural Information Processing Systems*.
- [179] Chenlin Meng, Ruiqi Gao, Diederik P Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. 2022. On Distillation of Guided Diffusion Models. In *NeurIPS 2022 Workshop on Score-Based Methods*.
- [180] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.
- [181] Chenlin Meng, Jiaming Song, Yang Song, Shengjia Zhao, and Stefano Ermon. 2020. Improved Autoregressive Modeling with Distribution Smoothing. In *International Conference on Learning Representations*.
- [182] Chenlin Meng, Jiaming Song, Yang Song, Shengjia Zhao, and Stefano Ermon. 2021. Improved Autoregressive Modeling with Distribution Smoothing. In *International Conference on Learning Representations*.
- [183] Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. 2021. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems* 34 (2021), 25359–25369.
- [184] Chenlin Meng, Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. 2020. Autoregressive score matching. *Advances in Neural Information Processing Systems* 33 (2020), 6673–6683.
- [185] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and Optimizing LSTM Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=SyyGPP0TZ>
- [186] Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association* 44, 247 (1949), 335–341.
- [187] Jiquan Ngiam, Zhenghao Chen, Pang W Koh, and Andrew Y Ng. 2011. Learning deep energy models. In *International Conference on Machine Learning*. 1105–1112.
- [188] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*. 8162–8171.
- [189] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. 16784–16804.
- [190] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion Models for Adversarial Purification. *arXiv preprint arXiv:2205.07460* (2022).

- [191] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. 2019. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. *arXiv preprint arXiv:1903.12370* (2019).
- [192] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. 2019. On Learning Non-Convergent Short-Run MCMC Toward Energy-Based Model. *arXiv preprint arXiv:1904.09770* (2019).
- [193] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. 2020. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4474–4484.
- [194] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
- [195] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*.
- [196] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*.
- [197] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- [198] Muzaffer Özbeş, Salman UH Dar, Hasan A Bedel, Onat Dalmaç, Şaban Öztürk, Alper Güngör, and Tolga Çukur. 2022. Unsupervised Medical Image Translation with Adversarial Diffusion Models. *arXiv preprint arXiv:2207.08208* (2022).
- [199] George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing Flows for Probabilistic Modeling and Inference. *J. Mach. Learn. Res.* 22, 57 (2021), 1–64.
- [200] Giorgio Parisi. 1981. Correlation functions and computer simulations. *Nuclear Physics B* 180, 3 (1981), 378–384.
- [201] Sung Woo Park, Kyungjae Lee, and Junseok Kwon. 2021. Neural Markov Controlled SDE: Stochastic Optimization for Continuous-Time Data. In *International Conference on Learning Representations*.
- [202] William Peebles and Saining Xie. 2022. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748* (2022).
- [203] Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. 2022. Towards performant and reliable undersampled MR reconstruction via diffusion model sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 623–633.
- [204] Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. 2022. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*. PMLR, 17644–17655.
- [205] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [206] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial latent autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*. 14104–14113.
- [207] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).
- [208] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*. 8599–8608.
- [209] Konpat Preechakul, Nattanat Chathee, Suttisak Wizadwongsu, and Supasorn Suwananakorn. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10619–10629.
- [210] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. *arXiv preprint arXiv:2303.09535* (2023).
- [211] Yixuan Qiu, Lingsong Zhang, and Xiao Wang. 2019. Unbiased Contrastive Divergence Algorithm for Training Energy-Based Latent Variable Models. In *International Conference on Learning Representations*.
- [212] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.
- [213] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. 8748–8763.

- [214] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [215] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [216] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
- [217] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. 8821–8831.
- [218] Martin Raphan and Eero P Simoncelli. 2007. Learning to be Bayesian without supervision. In *Advances in neural information processing systems*. 1145–1152.
- [219] Martin Raphan and Eero P Simoncelli. 2011. Least squares estimation without priors or supervision. *Neural computation* 23, 2 (2011), 374–420.
- [220] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *International Conference on Machine Learning*. 8857–8868.
- [221] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*. 8857–8868.
- [222] Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs M Bergmann, and Roland Vollgraf. 2020. Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows. In *International Conference on Learning Representations*.
- [223] Lillian J Ratliff, Samuel A Burden, and S Shankar Sastry. 2013. Characterization and computation of local Nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 917–924.
- [224] Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning*. 1530–1538.
- [225] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*. 1278–1286.
- [226] Benjamin Rhodes, Kai Xu, and Michael U Gutmann. 2020. Telescoping Density-Ratio Estimation. In *Advances in Neural Information Processing Systems*, Vol. 33. 4905–4916.
- [227] Oren Rippel and Ryan Prescott Adams. 2013. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125* (2013).
- [228] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10684–10695.
- [229] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *arXiv preprint arXiv:2208.12242* (2022).
- [230] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*. 1–10.
- [231] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).
- [232] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [233] Tim Salimans and Jonathan Ho. 2021. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*.
- [234] Tim Salimans and Jonathan Ho. 2021. Should EBMs model the energy or the score?. In *Energy Based Models Workshop-International Conference on Learning Representations*.
- [235] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. 2019. High-dimensional multivariate forecasting with low-rank gaussian copula processes. In *Advances in Neural Information Processing Systems*, Vol. 32.

- [236] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36, 3 (2020), 1181–1191.
- [237] Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. 2021. Step-unrolled Denoising Autoencoders for Text Generation. In *International Conference on Learning Representations*.
- [238] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.
- [239] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*. Springer, 146–157.
- [240] Arne Schniebing, Yuanqi Du, Charles Harris, Arian Jamash, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. 2022. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695* (2022).
- [241] Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. 2021. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*. 9558–9568.
- [242] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. 2020. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382* (2020).
- [243] Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. 2022. Conditional simulation using diffusion Schrödinger bridges. *arXiv preprint arXiv:2202.13460* (2022).
- [244] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. 2012. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*. IEEE, 245–248.
- [245] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
- [246] John Skilling. 1989. The eigenvalues of mega-dimensional matrices. In *Maximum Entropy and Bayesian Methods*. Springer, 455–466.
- [247] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. 2256–2265.
- [248] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*, Francis R. Bach and David M. Blei (Eds.). 2256–2265.
- [249] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- [250] Ki-Ung Song. 2022. Applying Regularized Schrödinger-Bridge-Based Stochastic Process in Generative Modeling. *arXiv preprint arXiv:2208.07131* (2022).
- [251] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. 2021. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, Vol. 34. 1415–1428.
- [252] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, Vol. 32.
- [253] Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, Vol. 33. 12438–12448.
- [254] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. 2019. Sliced Score Matching: A Scalable Approach to Density and Score Estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*. 204. <http://auai.org/uai2019/proceedings/papers/204.pdf>
- [255] Yang Song and Diederik P Kingma. 2021. How to train your energy-based models. *arXiv preprint arXiv:2101.03288* (2021).
- [256] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. 2021. Solving Inverse Problems in Medical Imaging with Score-Based Generative Models. In *International Conference on Learning Representations*.
- [257] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
- [258] James C Spall. 2012. Stochastic optimization. In *Handbook of computational statistics*. Springer, 173–201.
- [259] Jiachen Sun, Weili Nie, Zhiding Yu, Z Morley Mao, and Chaowei Xiao. 2022. PointDP: Diffusion-driven Purification against Adversarial Attacks on 3D Point Cloud Recognition. *arXiv preprint arXiv:2208.09801* (2022).

- [260] Jaesung Tae, Hyeongju Kim, and Taesu Kim. 2021. EdiTTS: Score-based Editing for Controllable Text-to-Speech. *arXiv preprint arXiv:2110.02584* (2021).
- [261] Huachun Tan, Guangdong Feng, Jianshuai Feng, Wuhong Wang, Yu-Jin Zhang, and Feng Li. 2013. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies* 28 (2013), 15–27.
- [262] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, Vol. 34. 24804–24816.
- [263] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).
- [264] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Veličković, and Michal Valko. 2021. Bootstrapped representation learning on graphs. *arXiv preprint arXiv:2102.06514* (2021).
- [265] Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844* (2015).
- [266] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [267] Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. 2023. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. In *International Conference on Learning Representations*.
- [268] Arash Vahdat, Karsten Kreis, and Jan Kautz. 2021. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, Vol. 34. 11287–11302.
- [269] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. 2022. UniTune: Text-Driven Image Editing by Fine Tuning an Image Generation Model on a Single Image. *arXiv preprint arXiv:2210.09477* (2022).
- [270] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *The 9th ISCA Speech Synthesis Workshop*.
- [271] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel Recurrent Neural Networks. In *International Conference on Machine Learning*. Maria-Florina Balcan and Kilian Q. Weinberger (Eds.). 1747–1756.
- [272] Pascal Vincent. 2011. A connection between score matching and denoising autoencoders. *Neural computation* 23, 7 (2011), 1661–1674.
- [273] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*. 1096–1103.
- [274] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. 2022. Guided Diffusion Model for Adversarial Purification. *arXiv preprint arXiv:2205.14969* (2022).
- [275] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022. Diffusion-GAN: Training GANs with Diffusion. *arXiv preprint arXiv:2206.02262* (2022).
- [276] Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. 2021. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*.
- [277] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. 2021. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802* (2021).
- [278] Antoine Wehenkel and Gilles Louppe. 2021. Diffusion priors in variational autoencoders. *arXiv preprint arXiv:2106.15671* (2021).
- [279] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).
- [280] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. 2022. Deblurring via stochastic refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*. 16293–16303.
- [281] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. *arXiv preprint arXiv:2303.04671* (2023).
- [282] Hao Wu, Jonas Köhler, and Frank Noe. 2020. Stochastic Normalizing Flows. In *Advances in Neural Information Processing Systems*, Vol. 33. 5933–5944.
- [283] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2212.11565* (2022).

- [284] Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. 2022. Diffusion-based Molecule Generation with Informative Prior Bridges.
- [285] Quanlin Wu, Hang Ye, and Yuntian Gu. 2022. Guided Diffusion Model for Adversarial Purification from Random Noise. *arXiv preprint arXiv:2206.10875* (2022).
- [286] Shoule Wu and Ziqiang Shi. 2021. ItôTTS and ItôWave: Linear Stochastic Differential Equation Is All You Need For Audio Generation. *arXiv e-prints* (2021), arXiv-2105.
- [287] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2020. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys (CSUR)* (2020).
- [288] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.
- [289] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. 2022. AnoDDPM: Anomaly Detection With Denoising Diffusion Probabilistic Models Using Simplex Noise. In *IEEE Conference on Computer Vision and Pattern Recognition*. 650–656.
- [290] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. 2021. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804* (2021).
- [291] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. 2016. A theory of generative convnet. In *International Conference on Machine Learning*. 2635–2644.
- [292] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. 2022. Vector Quantized Diffusion Model with CodeUnet for Text-to-Sign Pose Sequences Generation. *arXiv preprint arXiv:2208.09141* (2022).
- [293] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S Jaakkola. 2021. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. In *International Conference on Learning Representations*.
- [294] Yutong Xie and Quanzheng Li. 2022. Measurement-conditioned Denoising Diffusion Probabilistic Model for Under-sampled Medical Image Reconstruction. *arXiv preprint arXiv:2203.03623* (2022).
- [295] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- [296] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2022. Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models. *arXiv preprint arXiv:2212.14704* (2022).
- [297] Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. 2021. Self-supervised graph-level representation learning with local and global structure. In *International Conference on Machine Learning*. 11548–11558.
- [298] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2021. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*.
- [299] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. 2022. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. *arXiv preprint arXiv:2211.08332* (2022).
- [300] Tijin Yan, Hongwei Zhang, Tong Zhou, Yufeng Zhan, and Yuanqing Xia. 2021. ScoreGrad: Multivariate Probabilistic Time Series Forecasting with Continuous Energy-based Generative Models. *arXiv preprint arXiv:2106.10121* (2021).
- [301] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diffsound: Discrete Diffusion Model for Text-to-sound Generation. *arXiv preprint arXiv:2207.09983* (2022).
- [302] Jie Yang, Ruijie Xu, Zhiqian Qi, and Yong Shi. 2021. Visual anomaly detection for images: A survey. *arXiv preprint arXiv:2109.13157* (2021).
- [303] Kevin Yang and Dan Klein. 2021. FUDGE: Controlled Text Generation With Future Discriminators. (2021).
- [304] Ling Yang and Shenda Hong. 2022. Omni-Granular Ego-Semantic Propagation for Self-Supervised Graph Representation Learning. *arXiv preprint arXiv:2205.15746* (2022).
- [305] Ling Yang and Shenda Hong. 2022. Unsupervised Time-Series Representation Learning with Iterative Bilinear Temporal-Spectral Fusion. In *International Conference on Machine Learning*. 25038–25054.
- [306] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. 2022. Diffusion-Based Scene Graph to Image Generation with Masked Contrastive Pre-Training. *arXiv preprint arXiv:2211.11138* (2022).
- [307] Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. 2020. Dpgn: Distribution propagation graph network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 13390–13399.

- [308] Ling Yang, Zhilong Zhang, Wentao Zhang, and Shenda Hong. 2023. Score-Based Graph Generative Modeling with Self-Guided Latent Diffusion. (2023). <https://openreview.net/forum?id=AykEgQNPJEK>
- [309] Ruihan Yang and Stephan Mandt. 2022. Lossy Image Compression with Conditional Diffusion Models. *arXiv preprint arXiv:2209.06950* (2022).
- [310] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. 2022. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481* (2022).
- [311] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: filling missing values in geo-sensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.
- [312] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. 2021. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*. 12062–12072.
- [313] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. 2019. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, Vol. 32.
- [314] Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. 2023. Diffusion Models and Semi-Supervised Learners Benefit Mutually with Few Labels. *arXiv preprint arXiv:2302.10586* (2023).
- [315] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).
- [316] Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. 2022. Latent Diffusion Energy-Based Model for Interpretable Text Modelling. In *International Conference on Machine Learning*. 25702–25720.
- [317] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. 2022. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571* (2022).
- [318] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432* (2021).
- [319] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. 2023. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*.
- [320] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. In *Advances in Neural Information Processing Systems*.
- [321] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).
- [322] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).
- [323] Qinsheng Zhang and Yongxin Chen. 2021. Diffusion Normalizing Flow. In *Advances in Neural Information Processing Systems*, Vol. 34. 16280–16291.
- [324] Qinsheng Zhang and Yongxin Chen. 2022. Fast Sampling of Diffusion Models with Exponential Integrator. *arXiv preprint arXiv:2204.13902* (2022).
- [325] Qinsheng Zhang, Molei Tao, and Yongxin Chen. 2022. gDDIM: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564* (2022).
- [326] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
- [327] Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. 2022. Cross Reconstruction Transformer for Self-Supervised Time Series Representation Learning. *arXiv preprint arXiv:2205.09928* (2022).
- [328] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* (2016).
- [329] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. 2022. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635* (2022).
- [330] Yue Zhao, Zain Nasrullah, and Zheng Li. 2019. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research* 20 (2019), 1–7.

- [331] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022. Truncated diffusion probabilistic models. *arXiv preprint arXiv:2202.09671* (2022).
- [332] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations*.
- [333] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81.
- [334] Linqi Zhou, Yilun Du, and Jiajun Wu. 2021. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5826–5835.
- [335] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. 2022. Discrete contrastive diffusion for cross-modal and conditional generation. *arXiv preprint arXiv:2206.07771* (2022).
- [336] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131* (2020).
- [337] Roland S Zimmermann, Lukas Schott, Yang Song, Benjamin A Dunn, and David A Klindt. 2021. Score-based generative classifiers. *arXiv preprint arXiv:2110.00473* (2021).