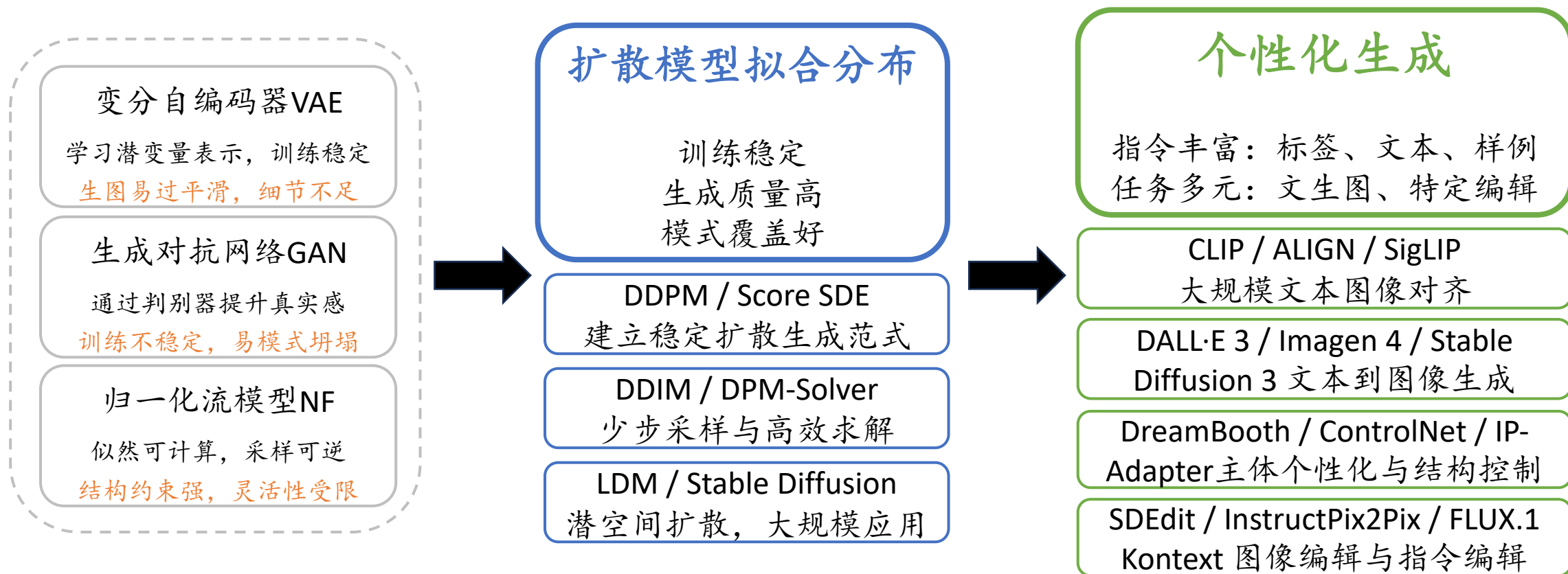


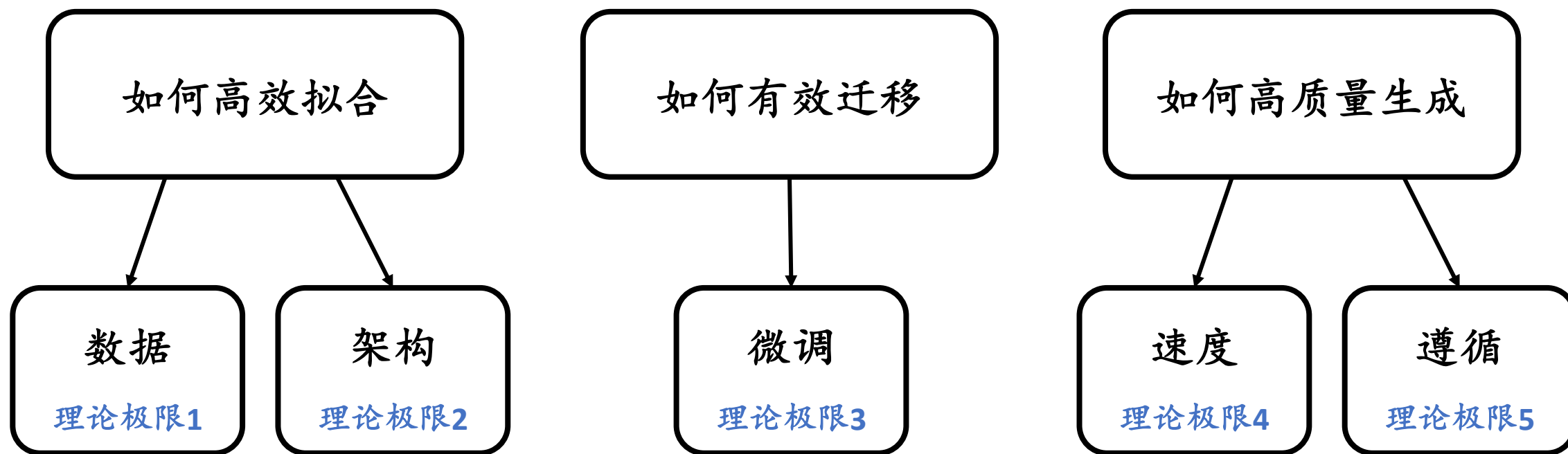
扩散模型的理论基础

李帅
上海交通大学

扩散模型：从高效数据拟合到高质量个性化生成



扩散生成模型的核心问题



理论极限1: 扩散模型拟合何种数据结构

• 理论可以证明

数据结构假设	得分函数结构	拟合准确度	参数量
低秩高斯模型	VAE+Linear Score	MNIST 8% CIFAR 25% ImageNet 0.254	64 1025
低秩子空间	VAE+UNet	MNIST 96% CIFAR 86% ImageNet 0.304	272K 641K
低秩混合高斯模型	VAE+MOE	MNIST 89% CIFAR 85% ImageNet 0.293 96.4%	6.5K 2.4% 41K 6.4%

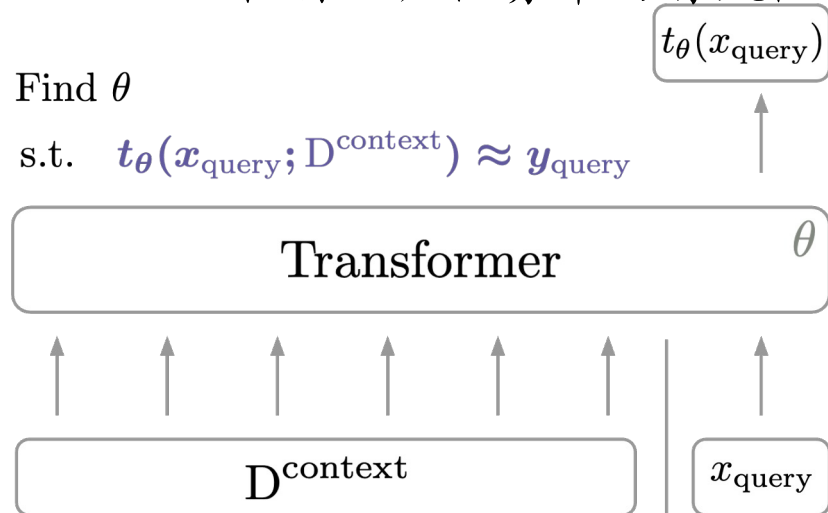
• 理论暂时无法证明

- 扩散模型拟合的核心数据具有什么结构？
- 准确率 VS 参数量
- 如果可以牺牲5%的准确率，最多可以节省多少参数量？
- 如果希望保留最多10%的参数量，准确率最高可以到多少？

理论极限2: Transformer为何在扩散模型中有效

- 理论可以证明

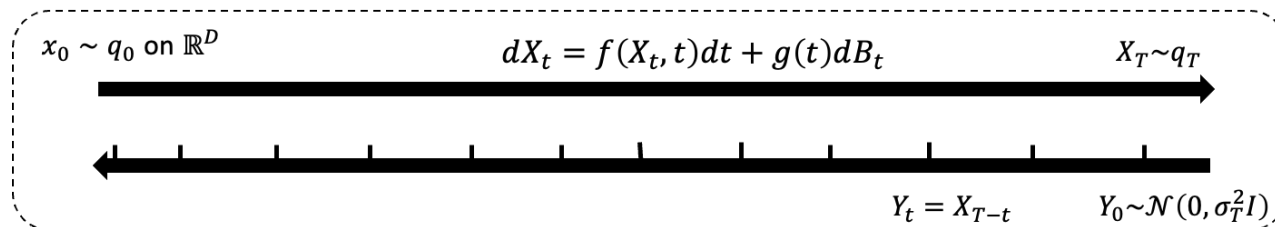
- Transformer在自回归任务中的有效性



- 定理. 存在Transformer可以实现

- Linear->Any Function Regression任务
- Linear->Any Function Classification任务
- Algorithm Selection
- 下的in-context learning
- $\Rightarrow$ 几乎所有的连续型任务

- 理论暂时无法证明



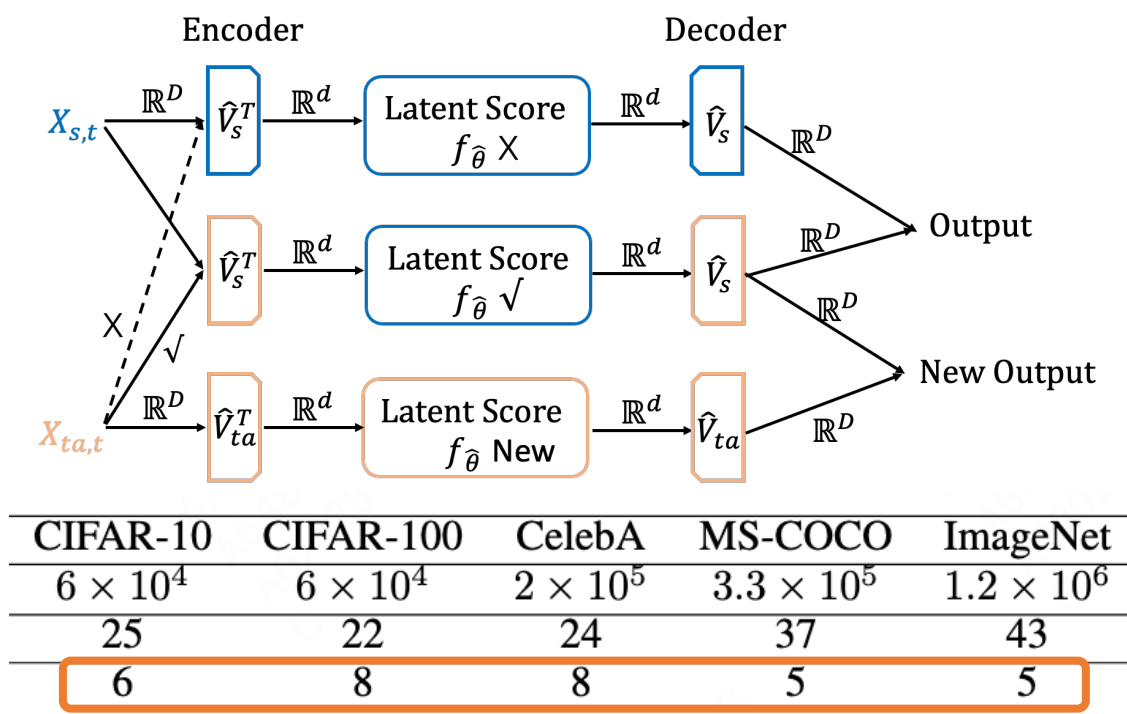
score function

$$dY_t = \left[-f(Y_t, T-t) + \frac{1+\eta^2}{2} g^2(T-t) \nabla \log q_{T-t}(Y_t) \right] dt + \eta g(T-t) dB_t, \eta \in [0,1]$$

- 得分函数学习了扩散过程如何去噪
- 为何适应Auto-regressive的Transformer也能使扩散模型表现良好 (DiT) ?
- Transformer架构具体学习了去噪过程中的何种关键性质?

理论极限3: 微调学习中的概念迁移

- 理论可以证明
 - 低秩混合高斯建模
需要假设共享概念与独立性
可基本对齐实际中需要的少量样本数



- 理论暂时无法证明
 - 相关性的概念如何迁移?
 - 概念/Score Function的迁移和domain adaptation的关联性
 - 在一般的迁移下, 扩散模型的few-shot (极少量样本) 为何有效?

理论极限4: 渐近采样生成与有限采样生成

- 理论可以证明

- 在得分函数充分准确的前提下
- 多步确定性采样器 + 朗之万校正器

RF（修正流模型）	$O(1/\epsilon_{TV}\epsilon_{W_2}^6)$
VP（方差保守模型）	$O(1/\epsilon_{TV}^1\epsilon_{W_2}^8)$
VE（方差爆炸模型）	$O(1/\epsilon_{TV}^2\epsilon_{W_2}^7)$

- 单步流模型复杂度：首次给出多项式离散化复杂度，并超越VP/VE模型一步结果

RF单步生成复杂度	$O(L_f/\epsilon_{W_2}^{3+1/a})$
VE单步生成复杂度	$O(L_f/\epsilon_{W_2}^{3+2/a})$
VP单步生成复杂度	$O(L_f/\epsilon_{W_2}^7)$

- 理论暂时无法证明

- 得分函数有较大误差时不同模型的生成效率比较
- 有限采样步数下，RF模型有效性的理论支撑
- 理论上单步生成复杂度，RF比VE好63%，实际表现约好27%，gap来自哪里
- 生成过程中的步长调度等操作如何找到理论支撑和指导？

理论极限5: 条件生成的指令遵循能力

• 理论可以证明

- 理想条件采样: 精确条件得分函数足以驱动反向过程从条件分布中采样
$$\nabla \log q_t(x|c) = \nabla \log q_t(x) + \nabla \log q_t(c|x)$$
- 需要放大Guidance方向
$$S_{\text{CFG}} = S_{\text{uncond}} + w * (S_{\text{cond}} - S_{\text{uncond}})$$
- 独立数据: 分类置信度随 w 增加收敛
VE&RF: $\mathcal{P}(x_T, c) \geq 1 - O(w^{-1}(\log w)^2)$
VP: $\mathcal{P}(x_T, c) \geq 1 - O(w^{-e^{-T}}(\log w)^{2e^{-T}})$
- 相关数据: w 过大容易产生曝光, 需要控制 $w(k)$ 上界

• 理论暂时无法证明

- 在ImageNet的实验中, VE的指令遵循能力比VP好约40.5%, 而其理论对应的提升约好59.4%
- gap出现在哪里
- 引导过大会过爆, 引导过小无法对齐
- 如何通过理论思想设计自适应的引导过程?

扩散模型的未来发展

一、理解并建模真实的物理世界

- 语言逻辑与视觉逻辑如何统一
- 高维数据如何高效训练
- 物理规律如何融入视频生成中

潜在
方案

- 理解物理世界数据结构特性（极限1）
- 何种神经网络架构更适用于表示真实的物理世界（极限2）
- 核心物理概念的表达与迁移（极限3）

二、多模态下的高效实时推理

- 低延时可控的高效推理挑战
- 已有的指令遵循能力无法适应复杂的多模态

潜在
方案

- 有限采样步骤的理论极限（极限4）
- 引导权重的动态设计以自适应指令遵循效果和多样性需求（极限5）

致谢!



李帅

- 上海交通大学 副教授
- 研究方向: 强化学习/机器学习理论
- <https://shuaili8.github.io/>

学生:



杨若峰



王至捷



李永灿

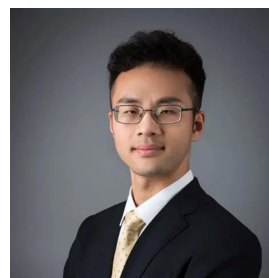


朱兆煜



仇翊羽

合作者:



江波



王昀翔



陈程



金睿楠

Questions?