《多模态大模型原理与应用》

Lecture 9
AIGC原理与应用

刘阳
中山大学
人机物智能融合实验室（HCP Lab）
liuy856@mail.sysu.edu.cn

什么是AIGC

AIGC（Artificial Intelligence Generated Content）是利用人工智能自动生成内容的技术体系

其生成范围覆盖：

- **文本：文章、代码、对话**
- **图像：绘画、设计、摄影**
- **视频：动画、影视、广告**
- **音频：音乐、语音、音效**
- **3D：模型、场景、资产**

AIGC在AI版图中的位置

AIGC是AI从"感知和理解"走向"创造和构造"的关键

传统AI任务 vs 生成式AI：

任务类型	传统AI	生成式AI
图像	分类、检测	生成新图像
文本	分类、翻译	写作、对话
音频	识别、转录	合成、作曲

AIGC可以生成什么

AIGC的生成对象已从文本扩展到图像、音频、视频、数字人和3D资产，是一类统一的生成问题

文本

文章、诗歌、代码、对话

图像

绘画、设计、摄影、海报

视频

动画、影视、广告片

音频

音乐、语音、配音、音效

3D资产

模型、场景、数字人

AIGC的典型应用场景

AIGC已广泛应用于多个领域，为后续原理学习提供动机

广告设计 — 快速生成创意素材与营销海报

教育内容 — 个性化教学材料与知识可视化

影视预演 — 概念设计与场景预览加速制作

游戏资产 — 角色、场景、纹理的自动化生成

电商营销 — 商品图、模特图、详情页批量产出

从判别式AI到生成式AI

判别式模型

回答"这是什么？"

- 图像分类
- 目标检测
- 语义分割
- 情感分析
- 问答系统

VS

生成式模型

回答"如何生成新的它？"

- 文本生成
- 图像生成
- 视频生成
- 音频生成
- 3D内容生成

整体知识地图

概念基础 → 代表模型 → 生成系统 → 前沿方法 → 综合案例

概念基础：生成模型定义、数据分布、采样、潜变量、条件生成

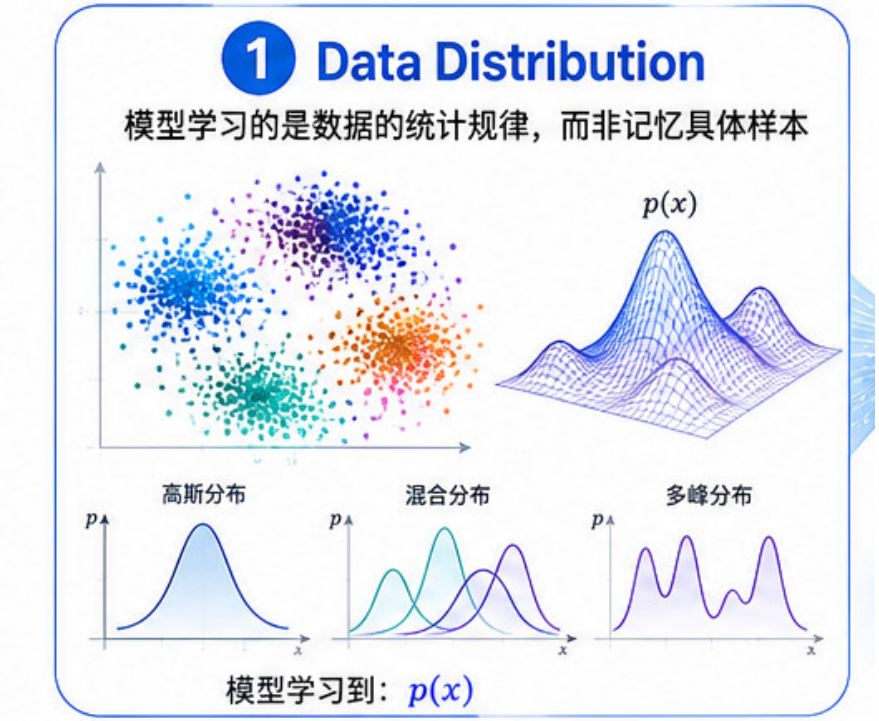
代表模型：自回归模型、VAE、GAN、扩散模型、流模型

生成系统：Latent Diffusion、文生图、图生图、ControlNet、LoRA

前沿方法：Flow Matching、Consistency Models、DiT、蒸馏加速

综合案例：视频生成、3D生成、多模态统一、Agent工作流

思考：机器为什么能创造



机器并非凭空创造，而是通过学习数据分布和条件关系生成新样本

关键视角：

- **数据分布：模型学习的是数据的统计规律，而非记忆具体样本**
- **条件生成：给定条件（如文本），模型从条件分布中采样**
- **组合创新：新样本是训练数据特征的重新组合**

带着这两个核心概念，进入后续内容的学习

案例：一句话生成一张海报

自然语言提示可以成为生成系统的高层控制接口。只需输入一句话，系统就能理解意图并生成对应的视觉内容。

这引出了一个核心问题：文本如何控制图像生成？

> *"一只穿着宇航服的柴犬，站在月球表面，地球在背景中升起，赛博朋克风格"*

本节内容

CONTENTS

一、生成模型基础

二、生成对抗网络GAN

三、扩散模型基础

四、视频、音频与3D生成

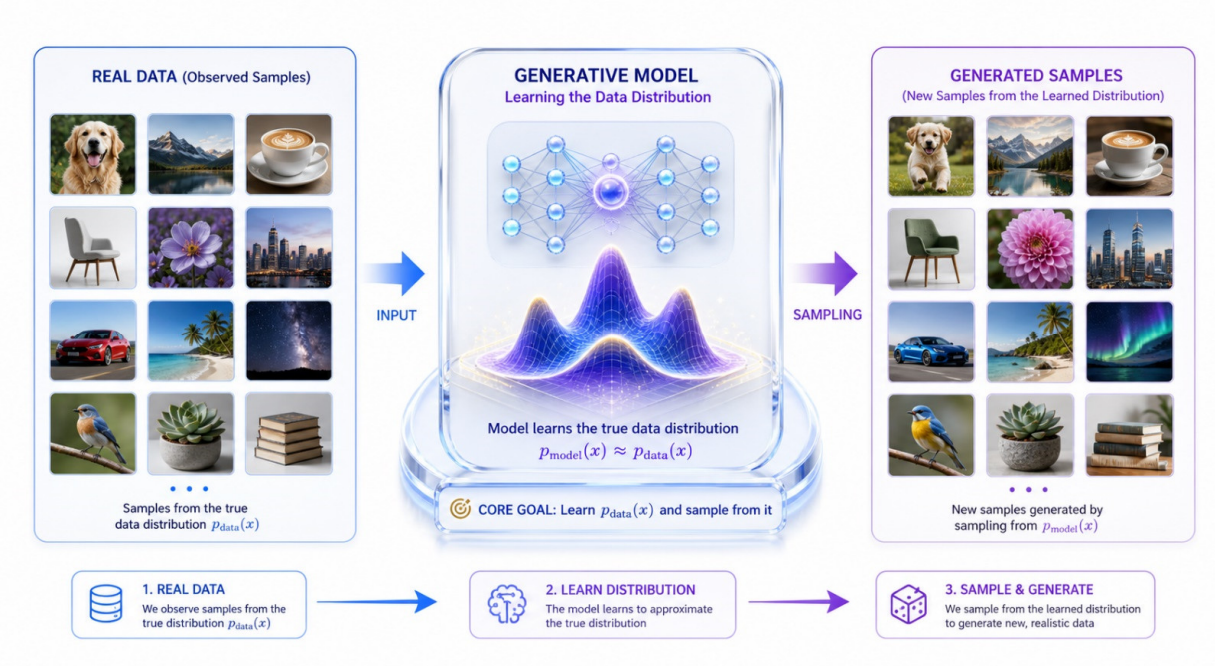
五、流匹配与DiT

什么是生成模型

生成模型的核心目标是学习真实数据分布，并从中采样得到新样本

关键理解：

- **学"分布"而不是"背样本"**
- **掌握数据的内在结构和规律**
- **能够创造出训练集中不存在的新内容**



核心公式：生成模型学习 $p_{data}(x)$，然后从中采样 $x \sim p_{model}(x)$，使得 $p_{model} \approx p_{data}$

数据分布的直觉理解

真实数据并非杂乱无章，而是分布在具有结构的高维空间中。

二维点云

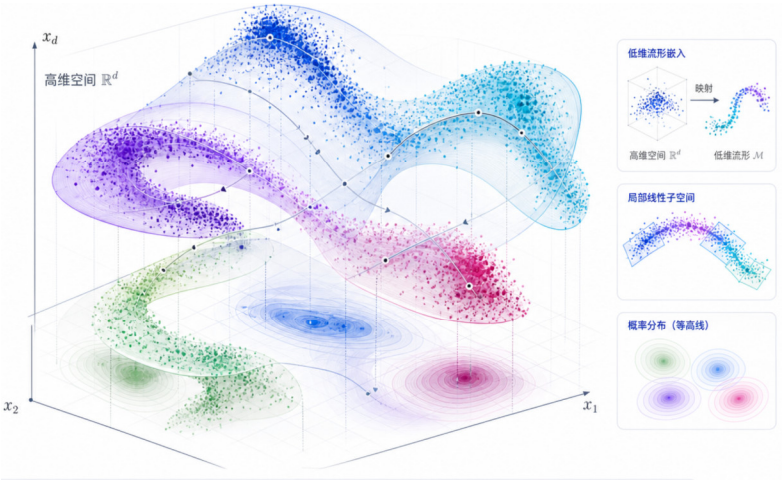
数据点在平面上形成特定形状（如环形、聚类），而非均匀散布

图像流形

所有"看起来像猫"的图像在高维空间中聚集成一个流形区域

结构约束

数据分布的"形状"决定了模型能生成什么样的有效内容

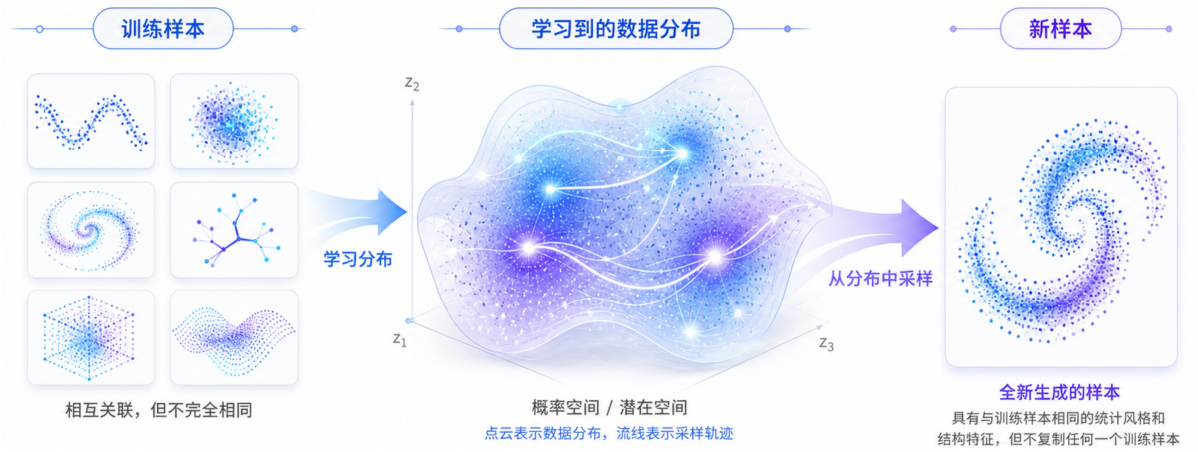


采样是什么

生成的本质不是"复制训练样本"，而是从学习到的分布中"采样出新内容"

生成 vs 检索：

- **检索：从已有数据中查找**
- **生成：创造全新的、合理的样本**



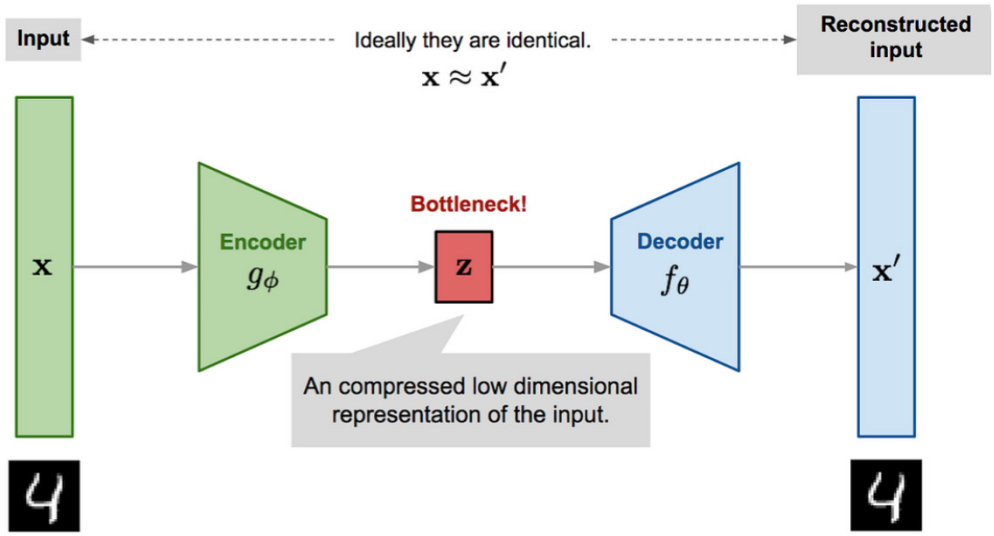
类比理解：就像画家学习了很多风景画的技法后，能画出从未见过的风景 — 不是复制某幅画，而是创造出符合"风景规律"的新作品

潜变量的概念

潜变量是控制内容变化的隐藏因素

例如在人脸图像中：

- **姿态：朝向、角度**
- **风格：写实、卡通**
- **光照：明暗、方向**
- **布局：构图、位置**



核心洞察：模型在隐空间中组织复杂内容 — 改变潜变量的值，就能控制生成内容的属性，这是可控生成的基础

条件生成是什么

条件生成是指模型在给定特定条件下生成内容，为文生图、图生图等方法做铺垫

文本条件
提示词描述

标签条件
类别标签

图像条件
参考图、草图

姿态条件
人体姿态骨架

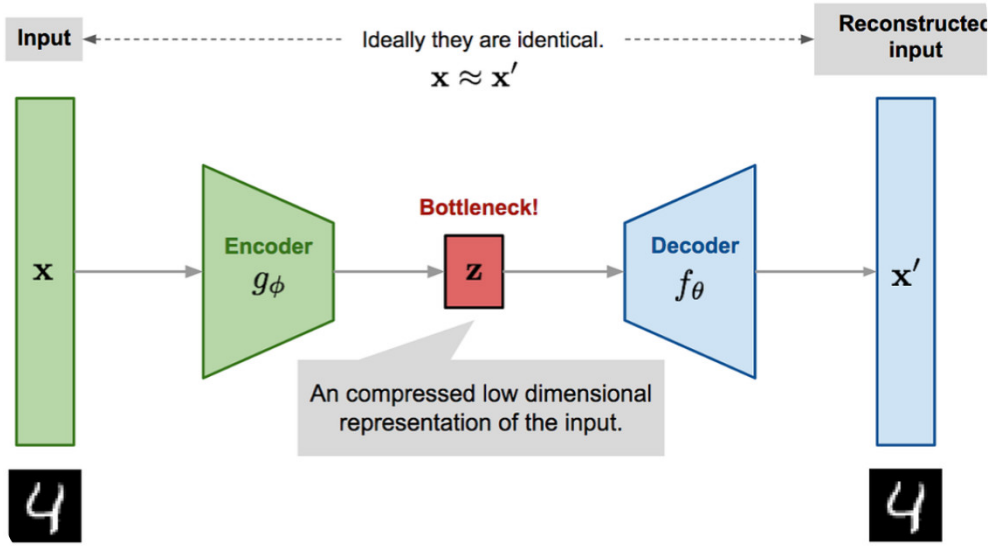
结构条件
边缘、深度图

生成模型的主要家族

常见生成模型的全景总览，建立完整的模型谱系

模型家族	核心思想	代表工作
自回归模型	逐元素预测序列	PixelCNN, GPT
VAE	编码-解码重建	VAE, β-VAE
GAN	生成器-判别器对抗	DCGAN, StyleGAN
扩散模型	逐步加噪再去噪	DDPM, Stable Diffusion
流模型	可逆神经网络变换	RealNVP, Glow

VAE的基本思想



VAE通过"编码到潜空间—再解码重建"学习数据生成过程

核心流程：

① 编码器：将输入映射到潜空间分布

② 采样：从潜分布中采样

③ 解码器：从潜变量重建输入

损失函数： $L = E_{q(z|x)}[\log p(x \mid z)] - D_{KL}(q(z \mid x) \| p(z))$ **— 重建损失 + KL散度正则化**

VAE的优点与局限

优点

- 训练稳定，有明确的损失函数
- 结构清晰，可解释性强
- 能学习有意义的潜表示
- 支持插值和编辑

局限

- 生成结果偏平滑
- 细节不够丰富
- 不如GAN逼真
- 存在后验坍塌风险

如何评价一个生成模型

从多个维度系统性地评价，而不只看个别"惊艳样例"

1 逼真度

生成样本与真实数据的相似程度

2 多样性

生成样本覆盖数据分布的广度

3 可控性

按指定条件生成目标内容的能力

4 稳定性

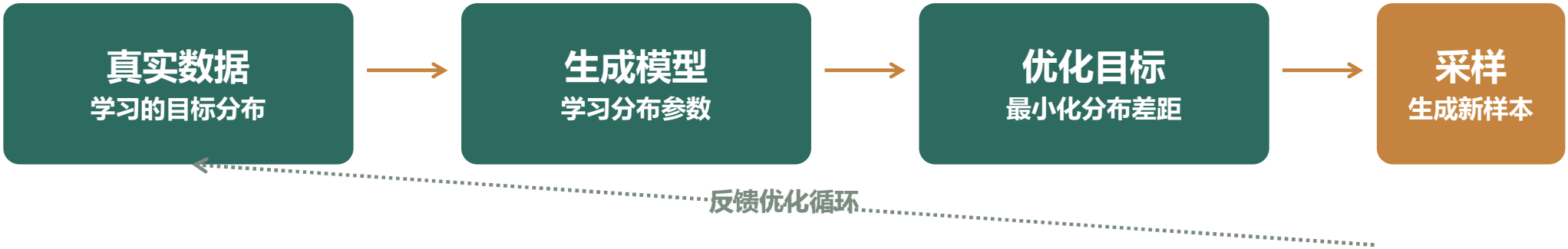
训练过程的收敛性和可复现性

5 效率

训练和推理的计算成本与速度

阶段总结：生成问题的统一框架

生成任务可统一理解为"数据—模型—目标—采样"的闭环过程



不同生成模型（VAE、GAN、扩散模型）的区别在于：如何定义模型结构、优化目标和采样策略，但本质上都遵循这一统一框架

本节内容

CONTENTS

GAN提出的背景

GAN的提出是为了**提升生成图像的真实感和细节质量**

GAN出现前的局限：

- 早期生成模型输出模糊
- 缺乏清晰的训练目标
- 难以生成高分辨率图像
- 视觉质量远不及真实图片

2014年，Ian Goodfellow 提出 GAN，开创了对抗训练的新范式。

其核心是：让两个网络相互竞争，比单个网络直接优化更能产生高质量结果

GAN的核心思想

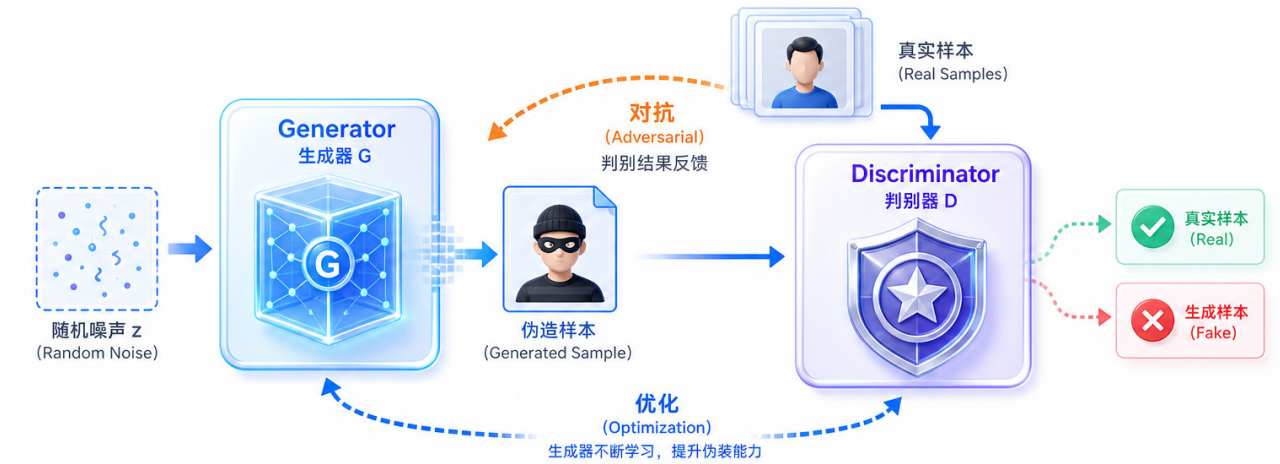
GAN通过生成器与判别器之间的对抗训练提升生成质量 — 就像"警察抓小偷"的博弈

生成器 G

- 像"伪造者"
- 目标是骗过判别器
- 将随机噪声变成假样本

判别器 D

- 像"鉴定专家"
- 目标是识别真假
- 为生成器提供训练信号



生成器在做什么

生成器学习把随机噪声映射成看起来像真实数据的样本

核心思想：

- "从简单分布到复杂分布"的映射
- 输入：简单噪声 $z \sim N(0, I)$
- 输出：复杂数据分布中的样本

生成器目标： $min_G \, E_{z \sim p_z(z)}[log(1 - D(G(z)))]$

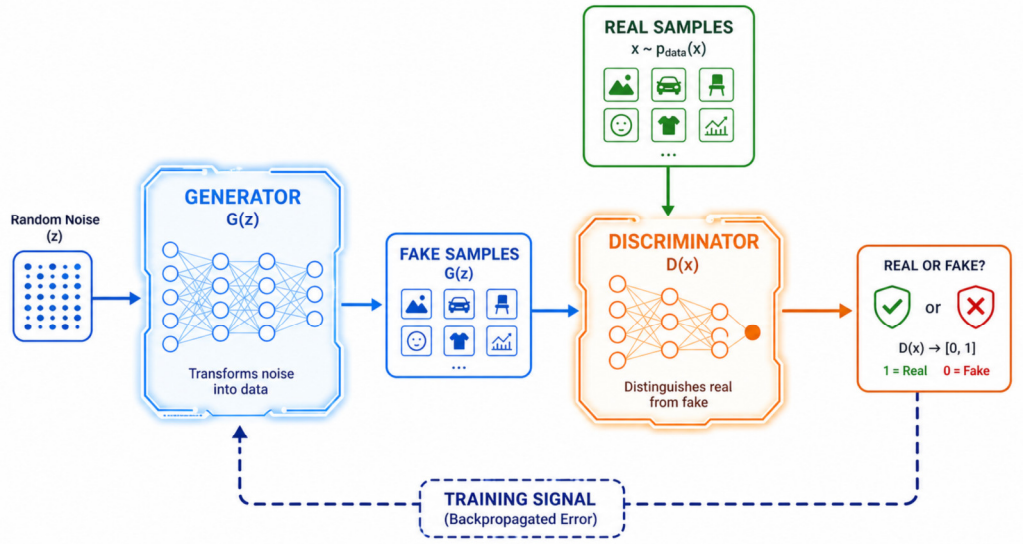
等价于让判别器对假样本的输出尽可能接近1（被判为真）

判别器在做什么

判别器负责区分真假样本，并为生成器提供训练信号

核心功能：

- 输入：真实样本或生成样本
- 输出：0（假）到 1（真）的概率
- 通过反向传播指导生成器改进

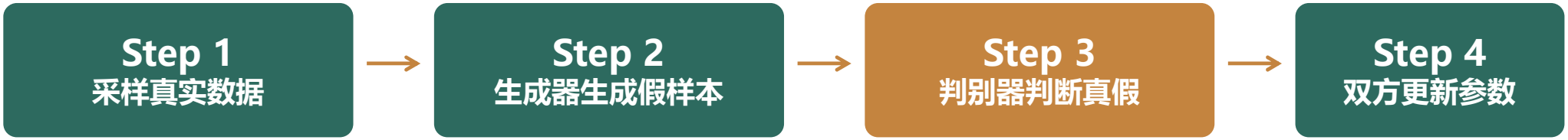


判别器目标： $max_D E_{x\sim p_{data}}[log\, D(x)] + E_{z\sim p_z}[log(1 - D(G(z)))]$

即：对真样本输出接近1，对假样本输出接近0

GAN的训练流程

GAN通过生成器和判别器交替优化实现动态博弈



关键要点：

- 每轮迭代中，判别器先更新 k 次，生成器更新 1 次
- 需要平衡双方能力：判别器太强 → 生成器梯度消失
- 判别器太弱 → 生成器得不到有效反馈
- 理想状态：纳什均衡，判别器无法区分真假

GAN为什么难训练

GAN易出现不稳定、梯度消失、震荡等问题 — "强对抗"既带来效果也带来困难

训练不稳定

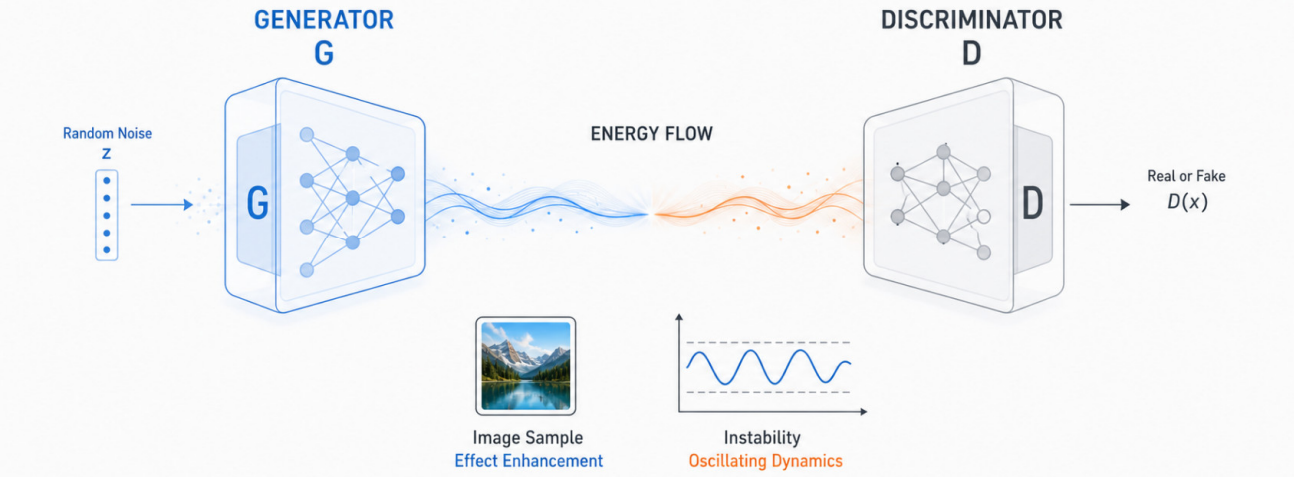
生成器和判别器的能力此消彼长，难以收敛

梯度消失

判别器过强时，生成器梯度极小，无法学习

损失震荡

损失函数不单调下降，难以判断训练进度

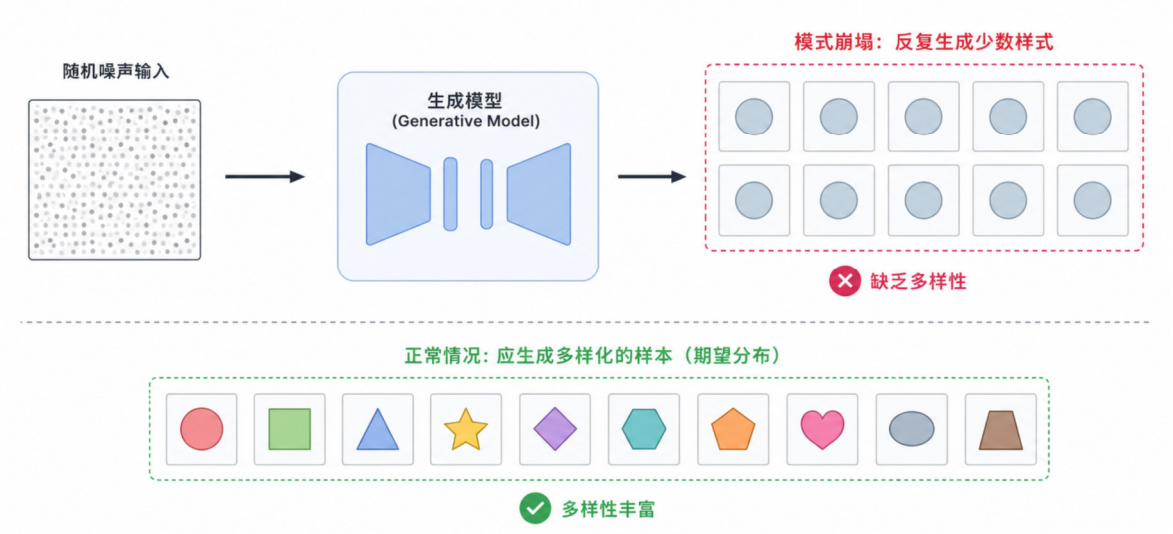


模式崩塌是什么

模式崩塌指模型反复生成少数样式，缺乏多样性

关键区分：

- "画得好看" ≠ "覆盖丰富"
- 生成器找到判别器的"盲点"
- 只生成最容易骗过判别器的样式



应对方法：

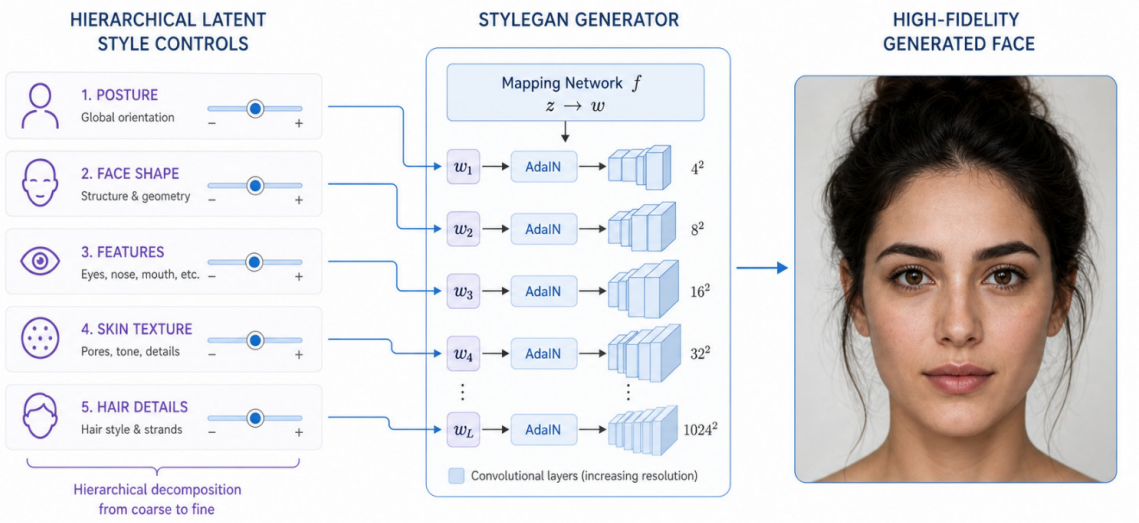
- Mini-batch discrimination：让判别器同时评估一批样本
- WGAN：用Wasserstein距离替代JS散度
- 经验技巧：标签平滑、噪声注入、历史样本缓冲

GAN的重要改进

GAN是一条持续演进的技术路线，关键改进工作如下

改进工作	改进方向	核心贡献
DCGAN	网络结构	提出稳定的卷积GAN架构
WGAN	损失函数	Wasserstein距离替代JS散度
LSGAN	损失函数	最小二乘损失替代交叉熵

StyleGAN的意义



StyleGAN通过分层控制潜在风格显著提升了高保真人脸生成能力

核心创新：

- 不同层控制不同尺度

低层→姿态、脸型
中层→ facial features
高层→颜色、纹理

关键洞察：StyleGAN将"风格"和"内容"解耦，使得用户可以独立控制不同层次的风格属性，实现了前所未有的可控人脸生成

GAN适合什么任务

GAN在高保真视觉任务中表现突出，总结其优势场景

人脸生成
StyleGAN系列的核心优势领域

头像/风格迁移
CycleGAN等跨域转换任务

图像增强
超分辨率、去噪、修复等任务

GAN的优势：单步生成速度快、视觉质量高、在结构化图像（人脸）上表现卓越

GAN的局限：训练困难、模式崩塌、可控性较弱 — 这些局限推动了扩散模型的崛起

本节内容

CONTENTS

一、生成模型基础

二、生成对抗网络GAN

三、扩散模型基础

四、视频、音频与3D生成

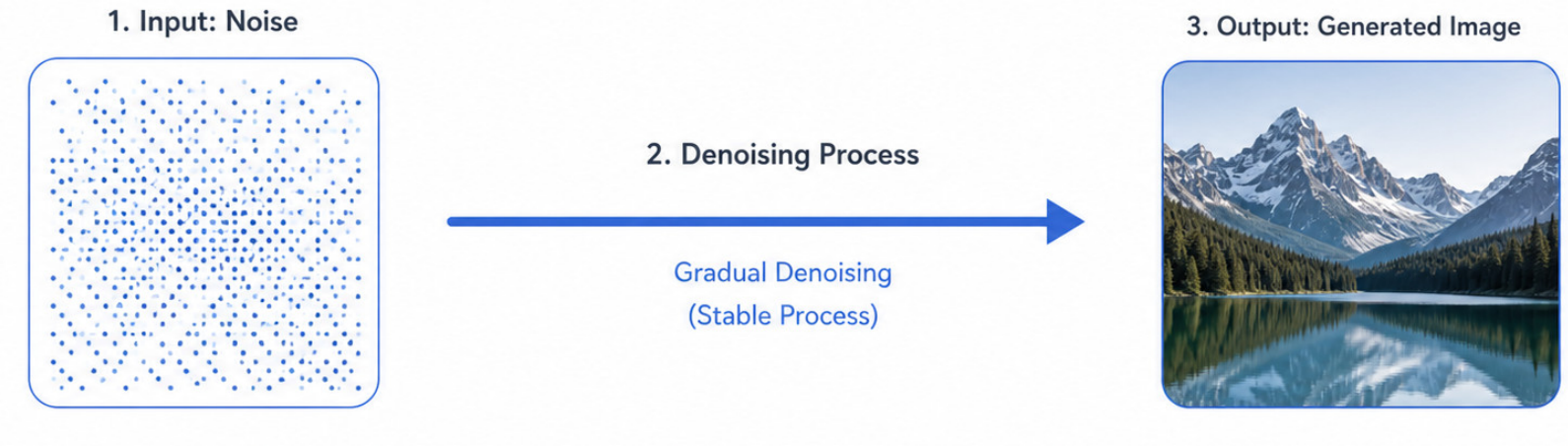
五、流匹配与DiT

为什么扩散模型会崛起

扩散模型以更稳定的训练方式实现了高质量、可控的生成能力

从GAN的局限自然过渡：

- **训练更稳定，不易崩溃**
- **生成质量持续提升**
- **天然支持条件生成**

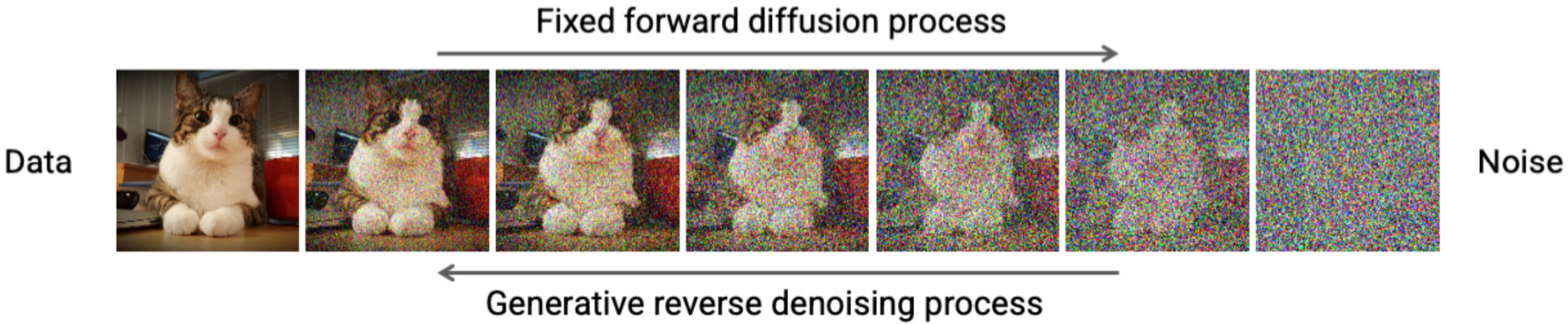


发展时间线：

2015 扩散概率模型提出 → 2020 DDPM突破 → 2022 Stable Diffusion 开源 → 2023 ControlNet/LoRA生态爆发

扩散模型的第一直觉

扩散模型先逐步把数据加噪，再学习如何逐步去噪恢复内容

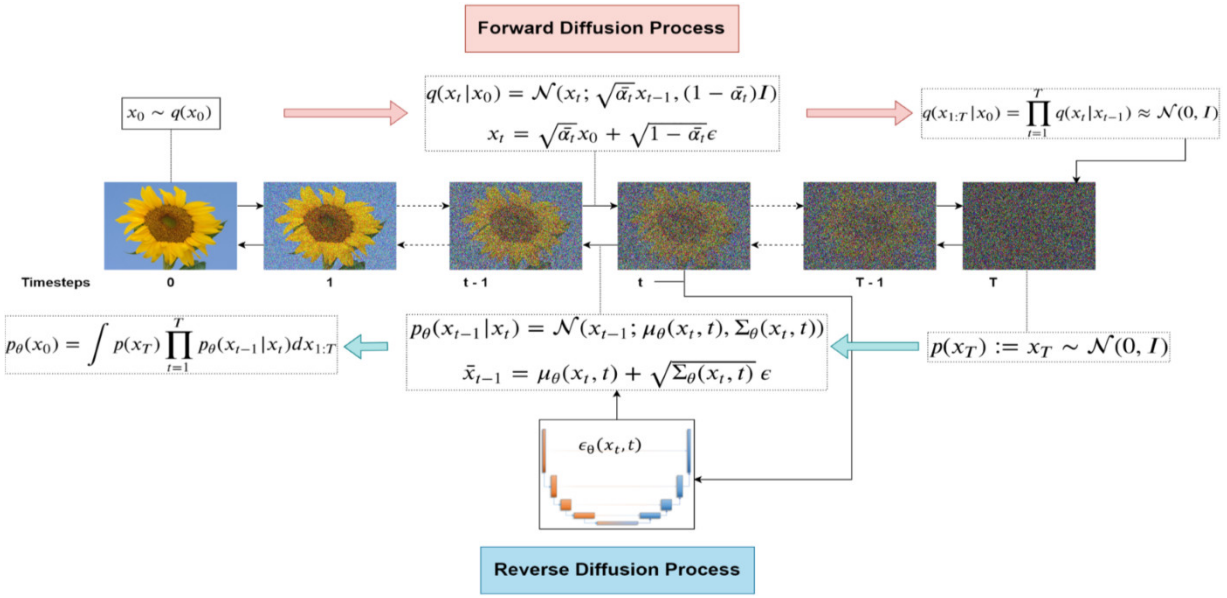


正向扩散过程

正向扩散通过不断注入小噪声，把原始样本逐渐变成近似纯噪声

特点：

- **过程是容易定义和可控的**
- **不需要学习，有闭式解**
- **每一步只加少量高斯噪声**



数学表达： $q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$

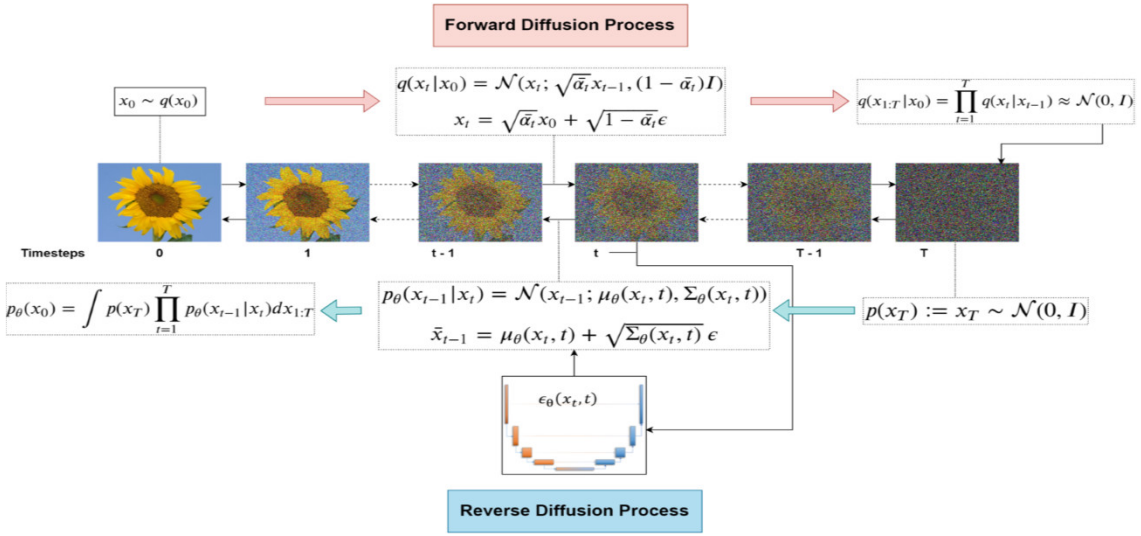
经过 T 步后， $x_T \approx N(0, I)$ **— 变成纯噪声**

反向去噪过程

反向过程学习如何从噪声中逐步恢复结构化内容

核心要点：

- 生成本质上是"逆扩散"
- 从纯噪声开始，逐步去噪
- 每一步预测并去除噪声



学习目标： $p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$

神经网络学习均值 μ_θ **和方差** Σ_θ**，指导去噪方向**

DDPM的基本框架

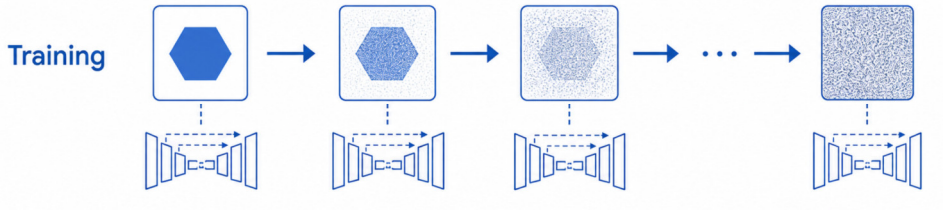
DDPM通过多步随机去噪实现高质量图像生成，包含训练和采样两条流程线

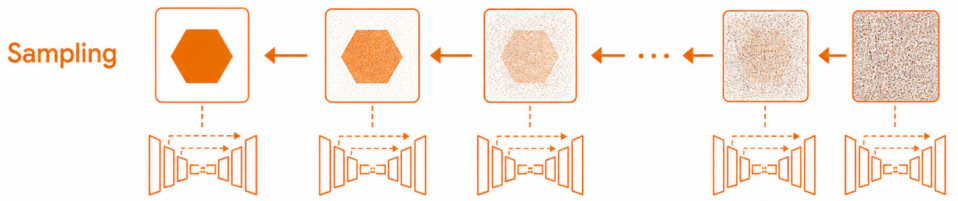
训练流程

① 从数据集中采样 x_0
② 随机选择时间步 t
③ 加噪得到 x_t
④ 网络预测噪声 $\epsilon_\theta(x_t, t)$
⑤ 最小化预测误差

采样流程

① 从纯噪声采样 x_T
② 从 $t = T$ 到 $t = 1$ 迭代
③ 每步预测噪声并去噪
④ 得到 x_{t-1}
⑤ 最终输出 x_0





模型究竟在预测什么

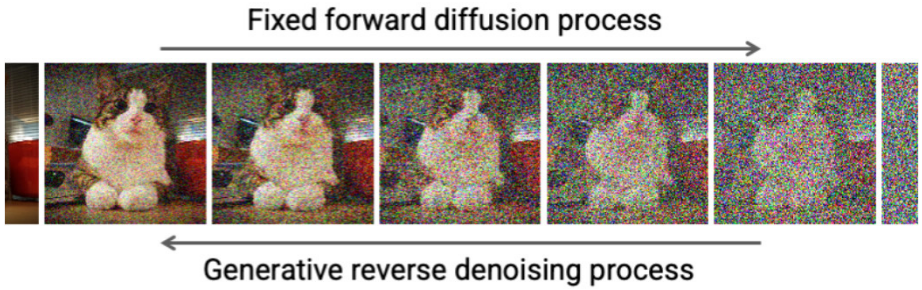
扩散模型通常不是直接预测最终图像，而是预测噪声、残差或速度

参数化方式	预测目标	代表模型
噪声预测	$\epsilon_\theta(x_t, t)$	DDPM, Stable Diffusion
样本预测	x_0 预测	部分变体
速度预测	v-parameterization	Salimans & Ho (2022)

为什么逐步去噪有效

把复杂生成任务拆成多个简单去噪步骤，

显著提升学习稳定性



工程优势：

- **每步只需做"一点点"修正**
- **比一步到位更容易学习**
- **训练过程更稳定可控**

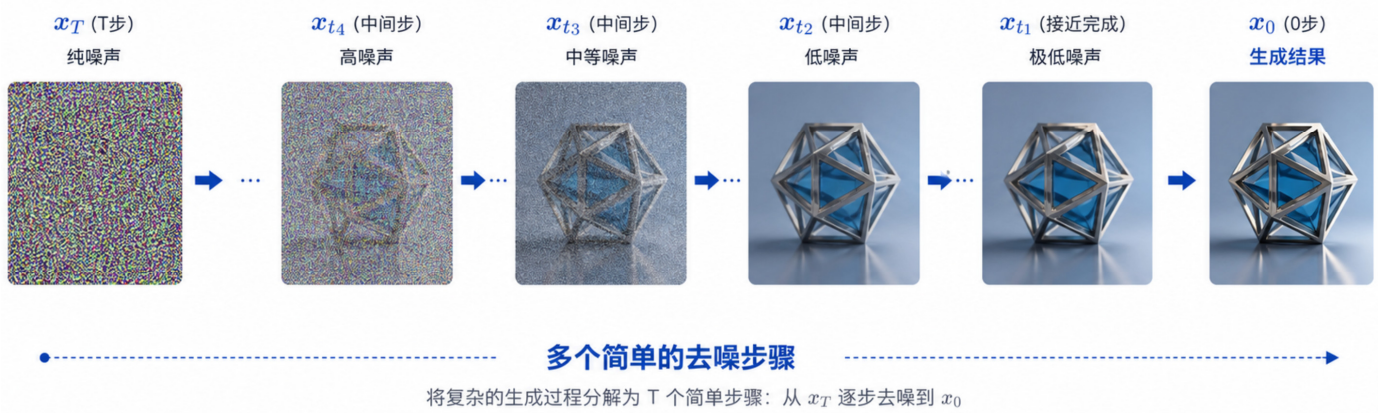
类比理解：就像从浓雾中逐步看清一幅画 — 每步只驱散一点点雾气，比一次性穿透浓雾容易得多。1000步去噪 ≈ 1000次小幅修正，累积成巨大的生成能力

扩散模型为什么慢

扩散模型常需多步迭代采样，推理成本较高

速度瓶颈：

- DDPM 需要 1000+ 步
- 每步都要过一遍网络
- 生成一张图需要数秒



加速方向：

- 采样加速：DDIM、DPM-Solver 减少步数到 20-50 步
- 模型蒸馏：将多步模型压缩为少步模型
- Consistency Models：一步生成
- Latent Diffusion：在压缩空间做扩散

噪声预测与样本预测

不同扩散模型可以预测噪声、原始数据或中间速度场，对应不同训练目标

噪声预测 (ϵ)

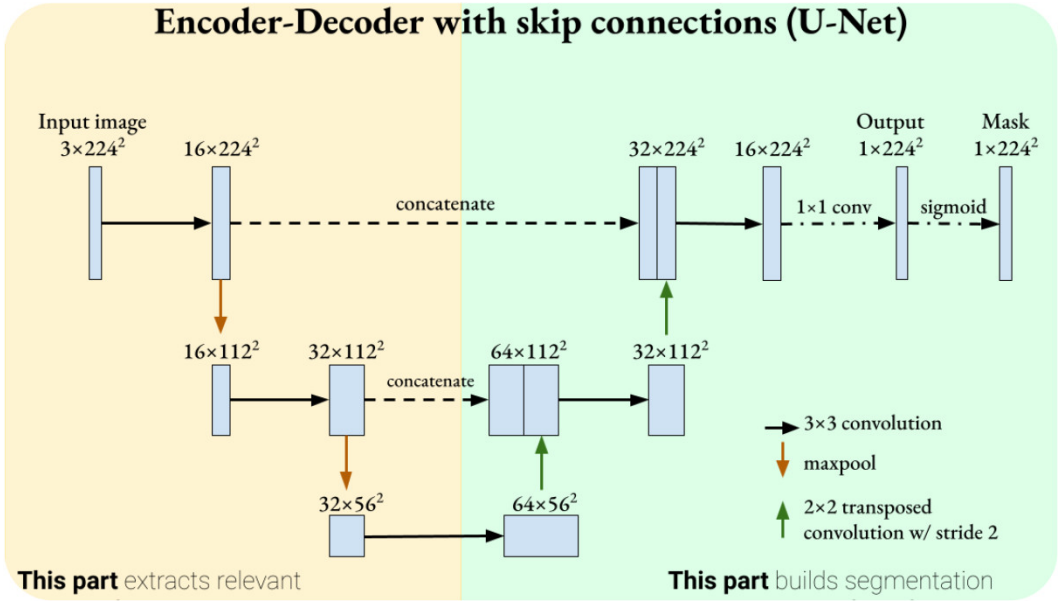
- 最常用
- 训练目标：MSE($\epsilon_\theta(x_t, t), \epsilon$)
- 直觉：猜当前加了什么噪声

v-预测 (Velocity)

- 更稳定的参数化
- 同时编码数据和噪声信息
- 在高分辨率生成中表现更好

参数化设计是扩散模型工程中的关键选择，不同的预测目标会影响训练稳定性、采样质量和速度

U-Net骨干网络



U-Net通过编码-解码和跳连接兼顾全局语义与局部细节

为什么扩散模型采用U-Net：

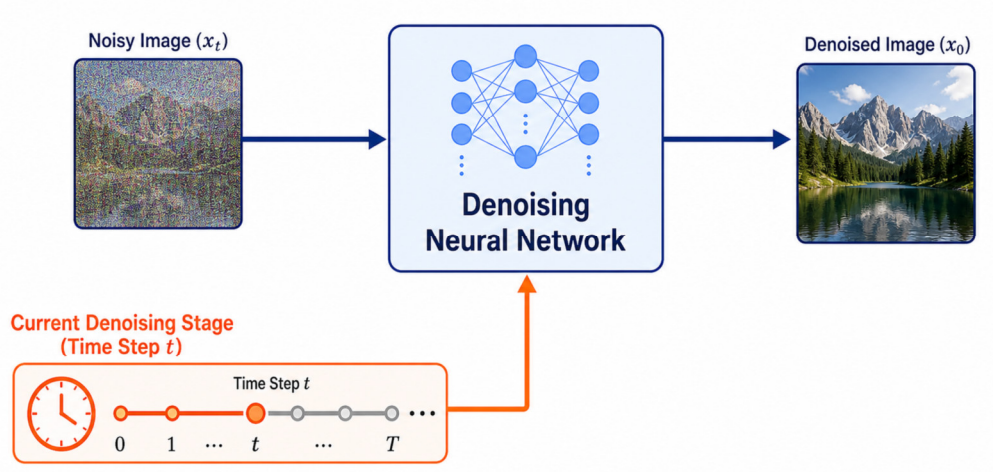
- 编码器提取多尺度特征
- 解码器恢复空间分辨率
- 跳连接保留细节信息
- 适合像素级预测任务

时间步嵌入

扩散模型需要显式知道当前处于哪一个去噪阶段

时间步编码的作用：

- 告诉网络"现在有多 noisy"
- 不同时间步需要不同处理
- 通常用正弦位置编码或学习嵌入



正弦位置编码： $PE(t)_{2i} = sin(t/10000^{2i/d})$, $PE(t)_{2i+1} = cos(t/10000^{2i/d})$

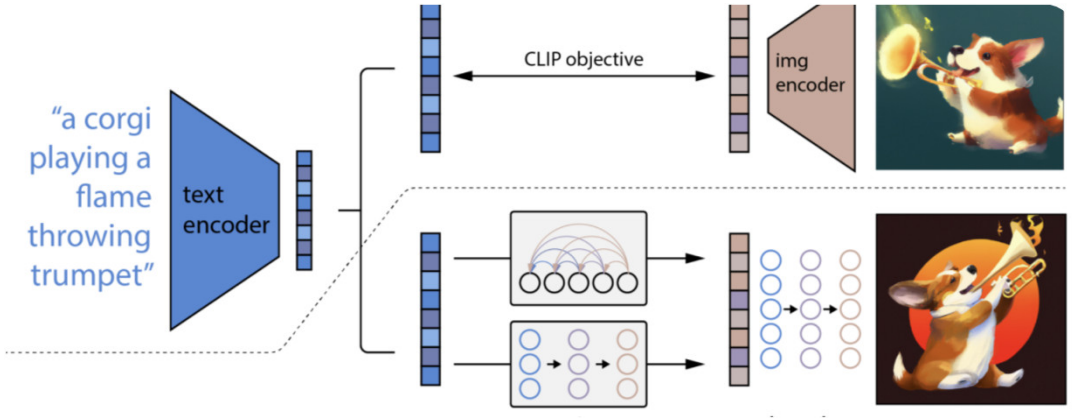
这种编码让网络能够区分不同时间步的噪声程度，是扩散网络的重要输入

条件扩散的概念

扩散模型可以在各种约束条件下进行条件生成

条件类型：

- 文本提示（Text Prompt）
- 类别标签（Class Label）
- 参考图像（Reference Image）



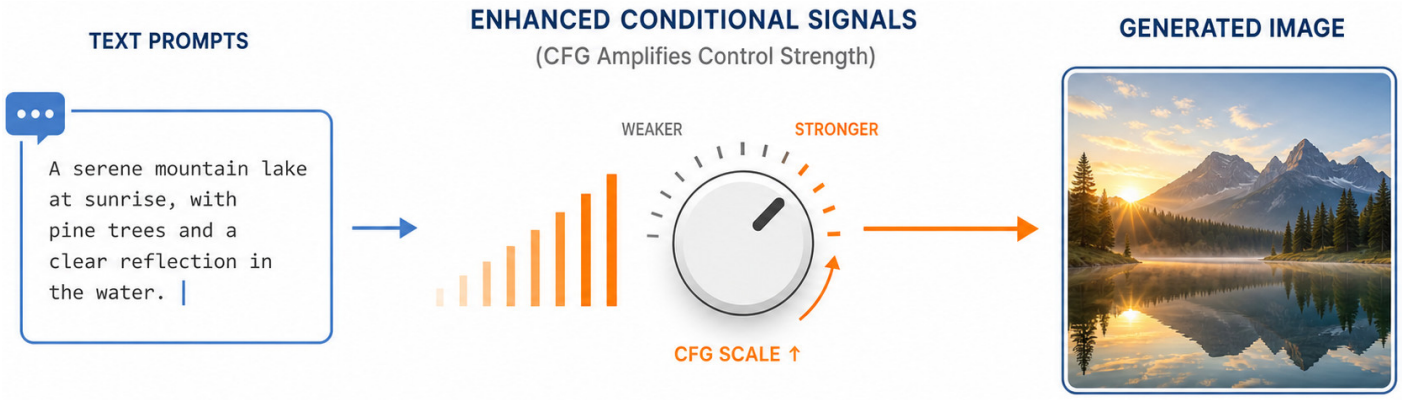
条件注入机制：

- 交叉注意力：将条件特征（如文本）与图像特征交互
- 特征拼接：将条件直接拼接到网络输入
- AdaGN/AdaIN：用条件调制归一化层参数

条件扩散是文生图系统的理论基础

Classifier-Free Guidance

CFG通过增强条件信号，让模型更"听话"地执行提示词



核心思想：

- 同时训练有条件和无条件生成
- 采样时放大条件影响
- Guidance Scale 是"控制强度旋钮"

公式：$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \emptyset) + s \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \emptyset))$

其中 s 为 guidance scale：s=1 无条件生成，s>1 增强条件遵循度

采样器与步数权衡

采样器的设计决定了生成速度与质量之间的平衡 — 同一模型因采样策略不同效果差异明显

采样器	典型步数	速度	特点
DDPM	1000	慢	最基础，质量最高
DDIM	50	中等	确定性采样，可跳步
DPM++	20	快	高阶求解器，质量损失小
Euler	20-30	快	简单高效，常用默认

阶段总结：扩散为什么成为主流

扩散模型兼具稳定性、可控性和多模态扩展能力，成为现代AIGC的重要基础

训练稳定

不易崩溃，收敛可靠

可控性强

CFG等机制精确控制

多模态扩展

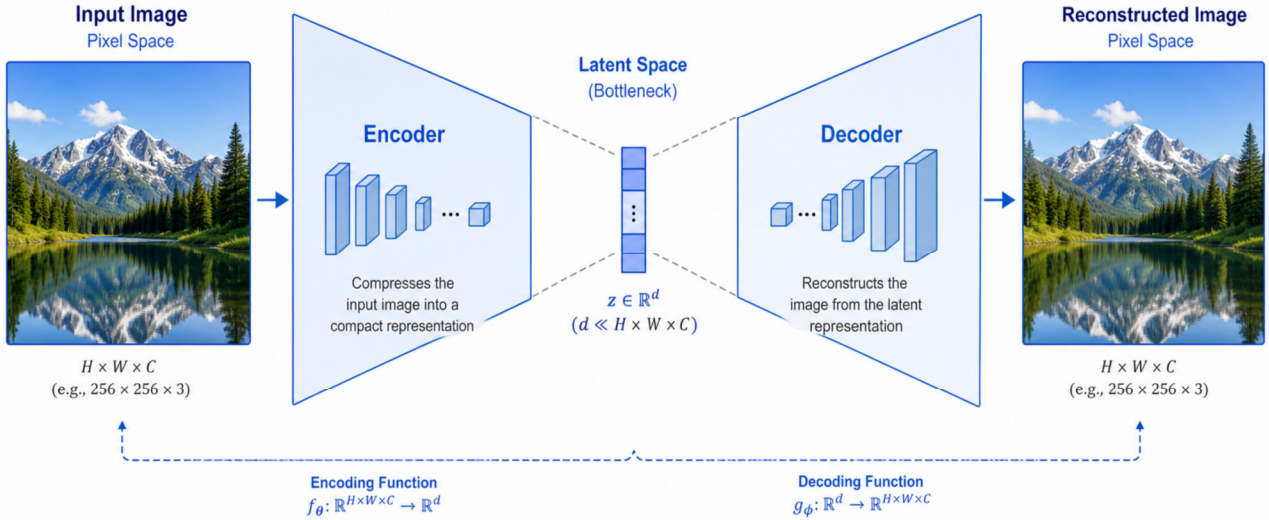
文本、图像、视频统一框架

潜空间扩散的思想

Latent Diffusion通过先压缩图像、再在潜空间做扩散，大幅降低计算成本

核心优势：

- 像素空间 → 低维潜空间
- 计算量减少数十倍
- 是Stable Diffusion的关键创新



流程：图像 $x \xrightarrow{Encoder} z \xrightarrow{扩散} z' \xrightarrow{Decoder} x'$

扩散过程在压缩后的潜空间 z 上进行，而非原始像素空间，大幅降低维度

编码器与解码器的作用

编码器把图像映射到低维潜空间，**解码器**把潜表示还原回像素空间

编码器 Encoder

- 输入：高分辨率图像
- 输出：低维潜表示
- 通常用VAE实现
- 压缩比约 1:48（SD）

解码器 Decoder

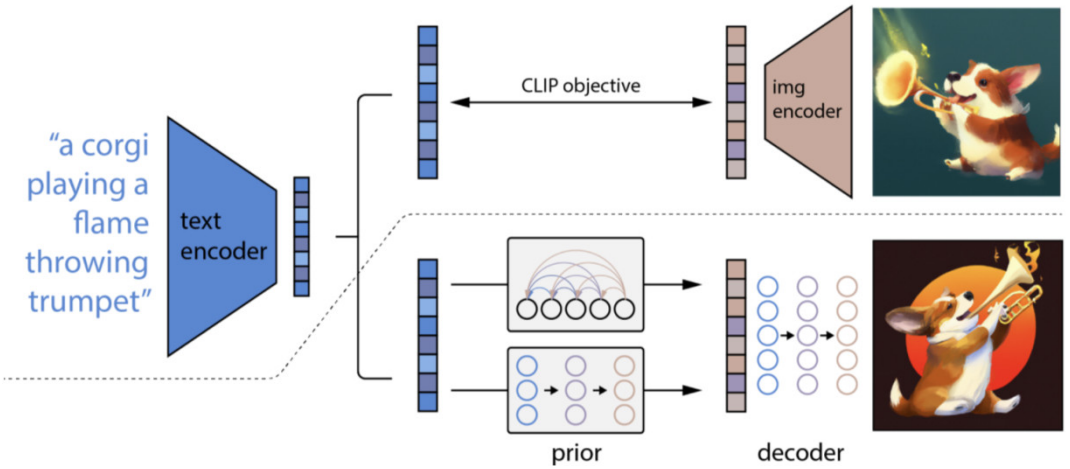
- 输入：潜表示
- 输出：重建图像
- 与编码器对称结构
- 负责最终像素还原

文本编码器如何工作

文本编码器把提示词变成语义向量，再注入扩散网络作为生成条件

工作流程：

- **文本 → Tokenizer 分词**
- **Token → Embedding 嵌入**
- **Transformer 编码语义**
- **输出文本特征向量**



常用文本编码器：CLIP Text Encoder（OpenAI）、T5（Google）、BERT变体

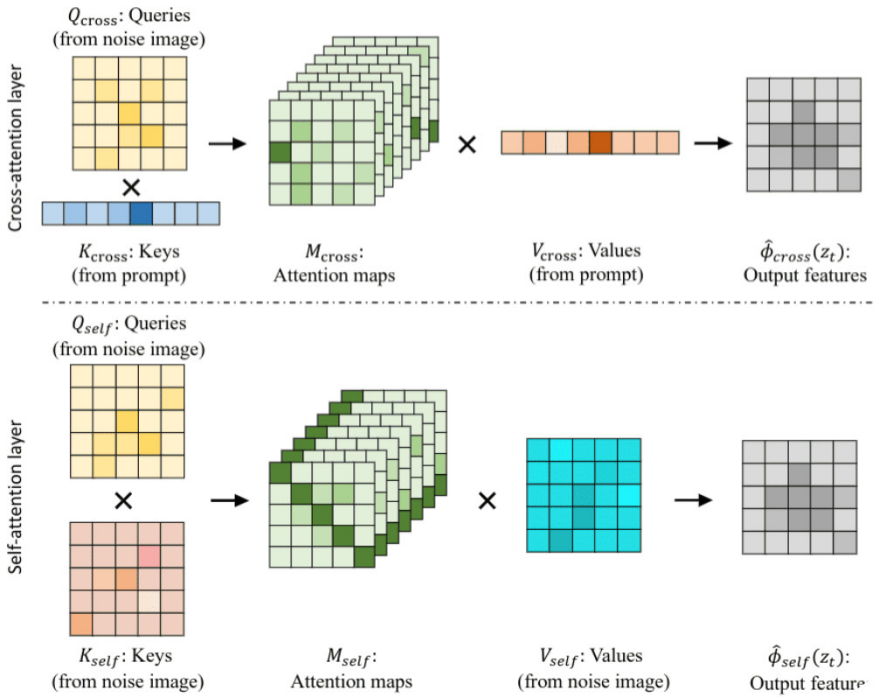
CLIP的优势：在图像-文本对上进行对比学习，生成的文本特征与视觉特征在同一语义空间中

交叉注意力机制

交叉注意力实现文本特征与图像潜表示之间的细粒度对齐

工作原理：

- **Query：来自图像特征**
- **Key/Value：来自文本特征**
- **图像"询问"文本该画什么**



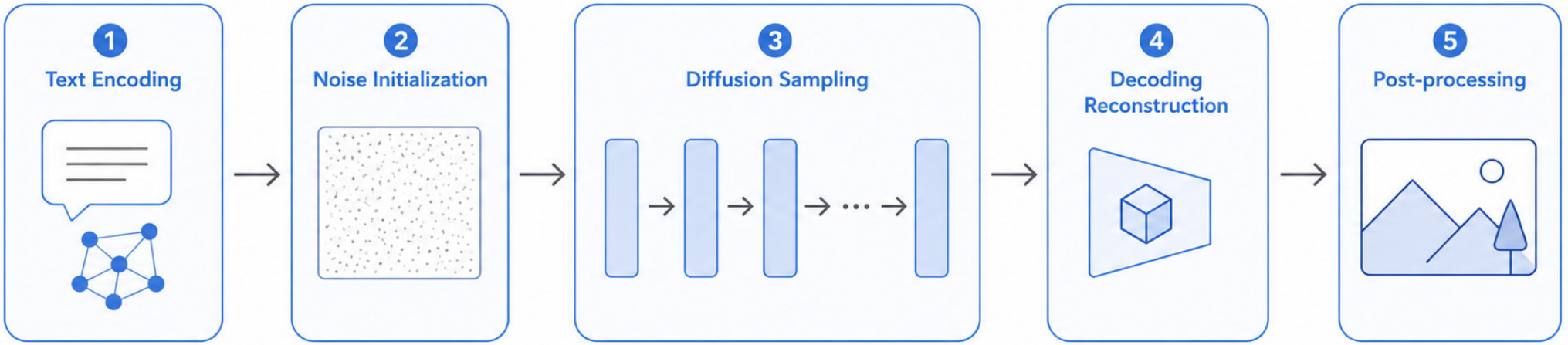
关键理解：交叉注意力是"文本控制图像生成"的核心机制

每个图像位置通过注意力权重，决定关注文本的哪些词，从而实现精确的语义对齐

文生图系统的完整流水线

典型文生图系统包括文本编码、噪声初始化、扩散采样、解码重建与后处理

文本编码
CLIP/T5
→
噪声初始化
随机潜噪声
→
扩散采样
U-Net去噪
→
解码重建
VAE Decoder
→
后处理

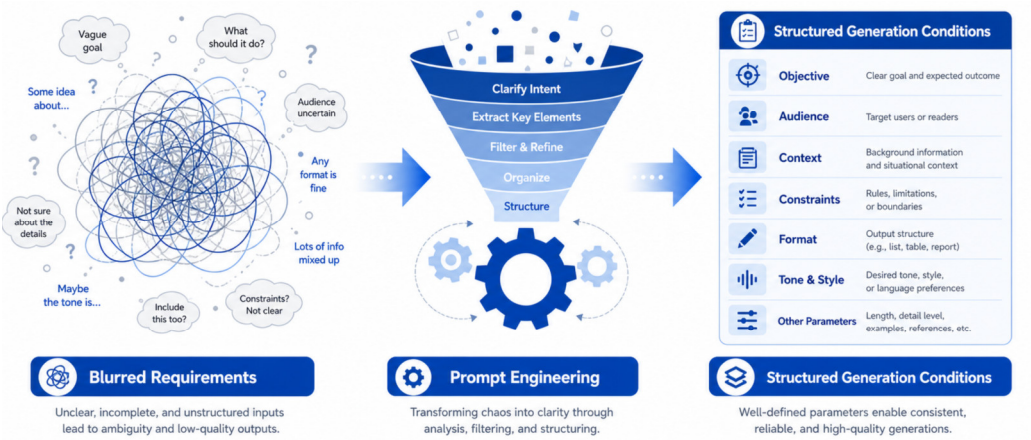


提示词工程的本质

提示词工程本质是把模糊需求变成结构化生成条件

核心理解：

- **提示词不是"玄学"**
- **而是任务表达方式**
- **好的提示词 = 清晰的指令**



类比：提示词就像给设计师的brief — 越具体、越结构化，输出越符合预期。模型不会"猜"你的意图，只会执行你表达的条件

正向提示词怎么写

一个好提示词通常包含主体、场景、风格、镜头、光线和细节描述

提示词构造模板：

[主体描述] + [场景/环境] + [艺术风格] + [镜头角度] + [光线条件] + [质量修饰词]

示例：

"一只橘猫（主体）坐在窗台上（场景），水彩画风格（风格），特写镜头（镜头），柔和的午后阳光（光线），高细节，8K（质量）"

负向提示词怎么用

负向提示词用于抑制瑕疵、错误结构和不希望出现的风格元素

生成控制是双向的：

- **正向："想要什么"**
- **负向："不要什么"**
- **两者结合实现精确控制**

常用负向提示词：

blurry, low quality, deformed, extra limbs, bad anatomy, watermark, text, signature

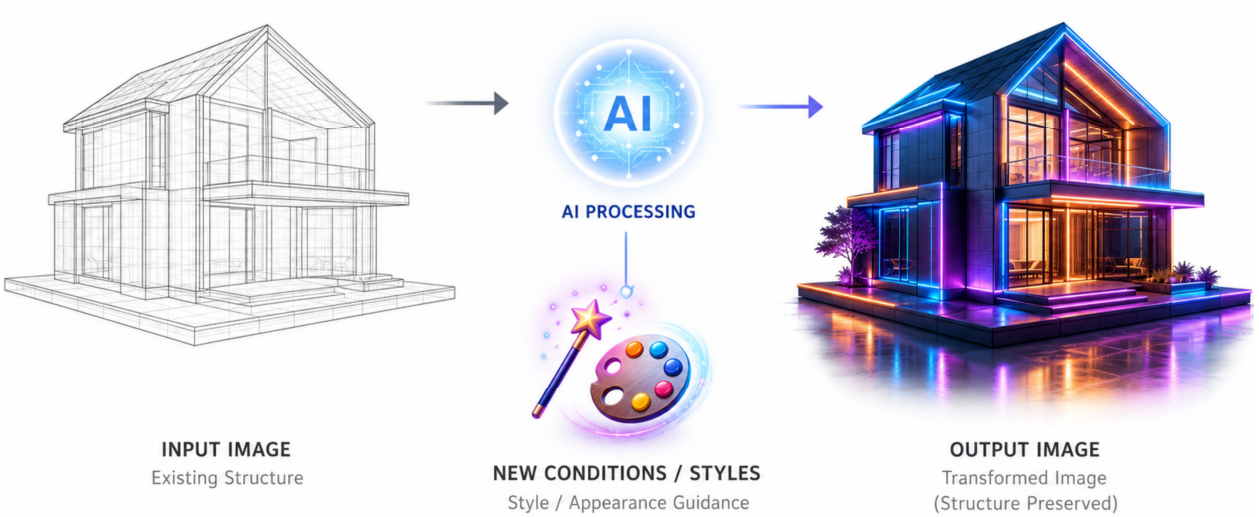
这些词告诉模型避免生成对应的不良特征

图生图的基本思路

图生图通过保留已有结构并引入新条件，实现内容变换和风格编辑

核心方法：

- 将输入图编码到潜空间
- 添加部分噪声（控制强度）
- 用新提示词引导去噪



Denoising Strength（去噪强度）控制保留原图结构的程度：

- 值越小（0.1-0.3）：保留更多原图结构
- 值越大（0.7-0.9）：更多创意变化

局部编辑 Inpainting

Inpainting通过掩码指定局部区域，实现局部重绘和精修

工作流程：

- 用户绘制掩码区域
- 掩码区域被重新生成
- 非掩码区域保持不变

应用价值："保持整体不变、只改局部"

- 移除不想要的物体
- 替换局部内容
- 修复图像缺陷

外扩编辑 Outpainting

Outpainting可以把已有图像向外扩展，补全更大场景

核心能力：

- 扩展画面边界
- 保持风格一致性
- 生成无缝衔接的新内容

应用场景：

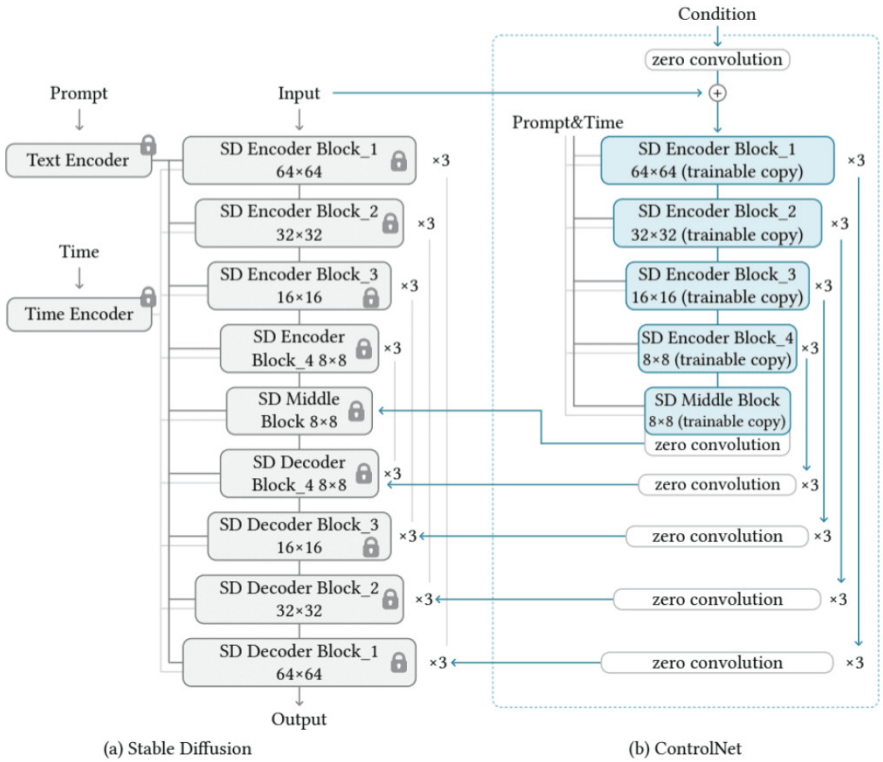
- 将竖构图扩展为横构图
- 补全画面边缘缺失的内容
- 创建全景图

ControlNet的核心思想

ControlNet通过额外控制支路把结构信息注入扩散过程

核心创新：

- 复制原U-Net的可训练副本
- 通过零卷积连接到主网络
- 不破坏原模型的生成能力



"可控生成"的重要方法：ControlNet让用户能精确控制生成图像的空间结构，是扩散模型工程化的里程碑

常见控制条件类型

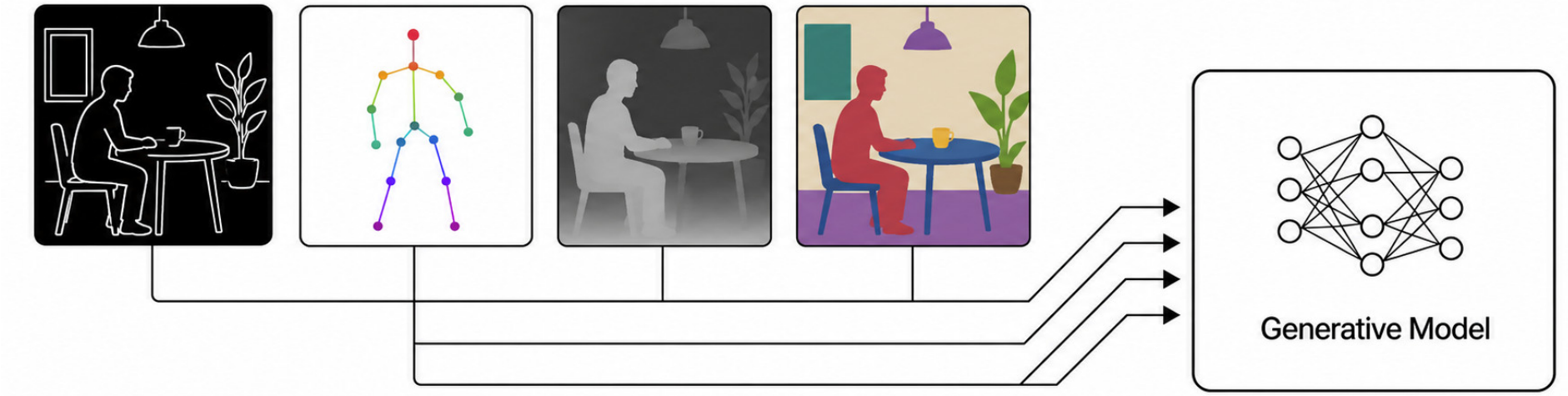
边缘图、人体姿态、深度图、语义分割图都可以成为生成条件

Canny边缘
控制轮廓和线条结构

OpenPose
人体姿态骨架控制

深度图
控制空间距离关系

语义分割
控制物体类别和布局

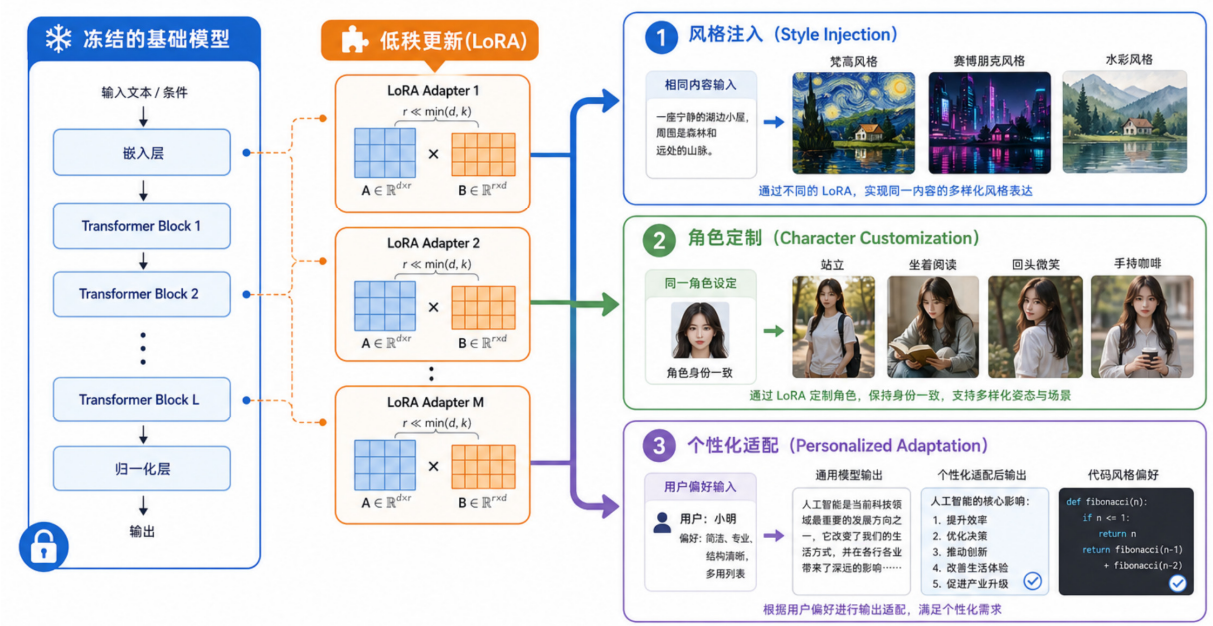


LoRA与轻量微调

LoRA通过低秩更新实现风格注入、角色定制和个性化适配

核心优势：

- **只训练少量参数（1%-5%）**
- **模型文件极小（几MB）**
- **可叠加多个LoRA使用**



原理：$W' = W + BA$，其中 $B \in R^{d\times r}, A \in R^{r\times k}$，秩 $r \ll min(d,k)$

这使得普通人也能训练自己的风格模型，推动了AIGC生态的繁荣

Textual Inversion与概念注入

Textual Inversion通过学习新的概念token扩展模型的表达能力

核心思想：

- "教会模型一个新概念"
- 不改变模型权重
- 只学习新的嵌入向量

应用场景：

- 让模型学会特定的艺术风格
- 教会模型认识特定角色或物体
- 用自定义token在提示词中调用新概念

本节内容

CONTENTS

一、生成模型基础

二、生成对抗网络GAN

三、扩散模型基础

四、视频、音频与3D生成

五、流匹配与DiT

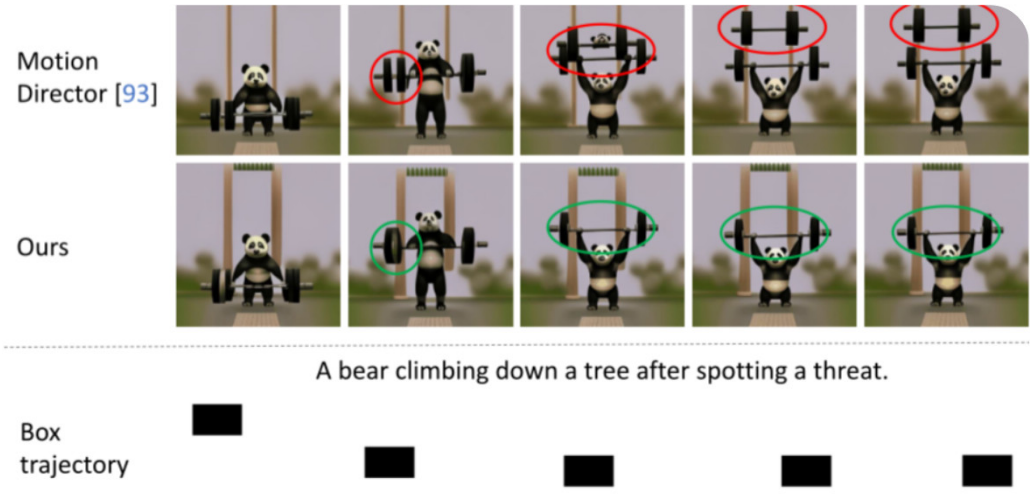
从图像生成走向视频生成

视频生成在图像质量之外，还必须解决

时间维度的一致性问题

核心挑战：

- **单帧质量 → 已相对成熟**
- **时间一致性 → 核心难点**
- **动作逻辑 → 更高层次要求**



关键区别：视频 = 图像 + 时间维度

生成视频不仅要"画得好"，还要"动得合理" — 这是从图像到视频的核心跨越

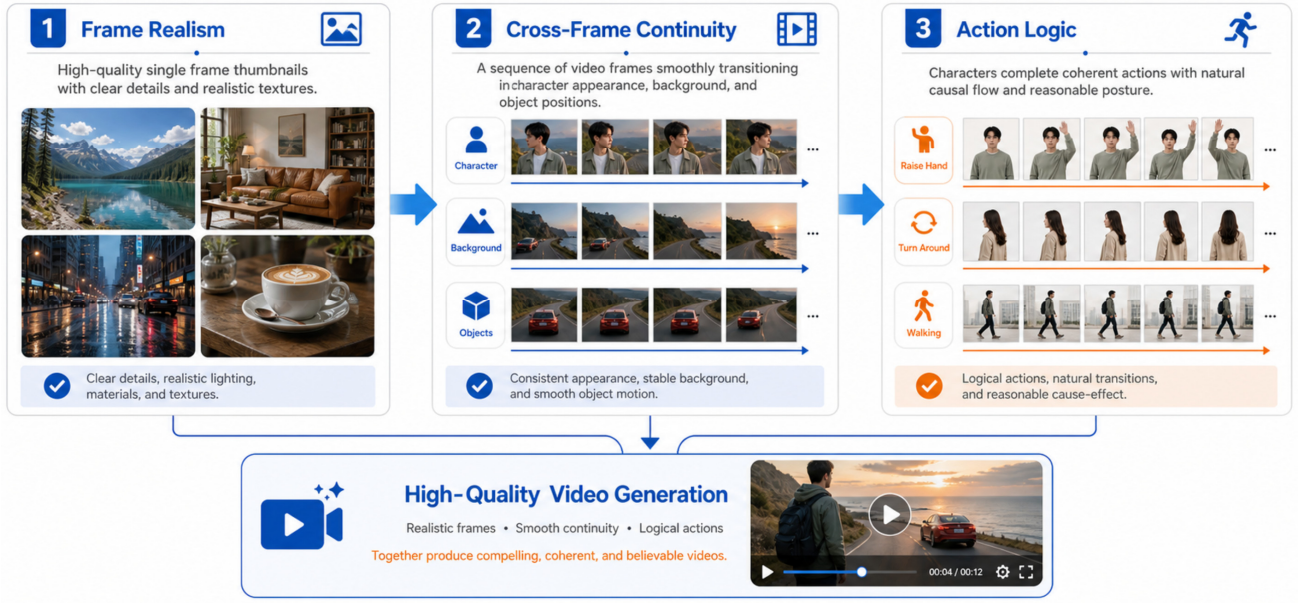
视频生成为什么更难

视频生成要同时保证单帧逼真、跨帧连续和动作逻辑合理

单帧逼真
每一帧都要像高质量图像

跨帧连续
帧与帧之间无闪烁、无跳变

动作逻辑
运动符合物理规律



时空一致性是什么

时空一致性指角色、背景、动作和物理过程在时间上的连续合理性

评价维度：

- 外观一致性：角色不变形
- 运动平滑性：无突兀跳变
- 物理合理性：符合自然规律

这是视频生成最关键的评价维度 — 即使单帧质量很高，如果时空一致性差，视频也会显得"假"和"不自然"

视频生成的主要技术路线

视频生成可采用逐帧建模、时空联合建模或图像到视频扩展等路线

逐帧建模

- 每帧独立生成
- 用条件保证连续性
- 简单但一致性差

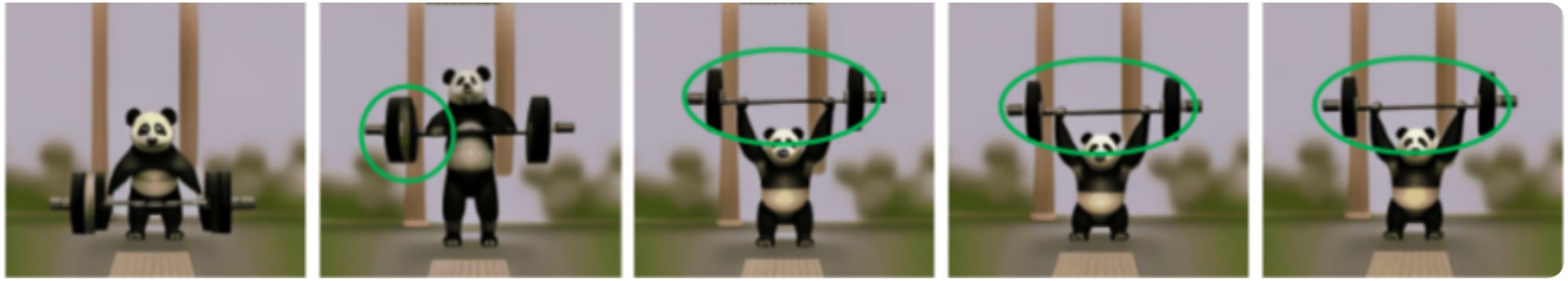
时空联合建模

- 3D卷积/注意力
- 同时建模时空
- 一致性好但计算大

图生视频扩展

- 以图像为起始帧
- 扩展为视频序列
- 当前主流方向

Ours

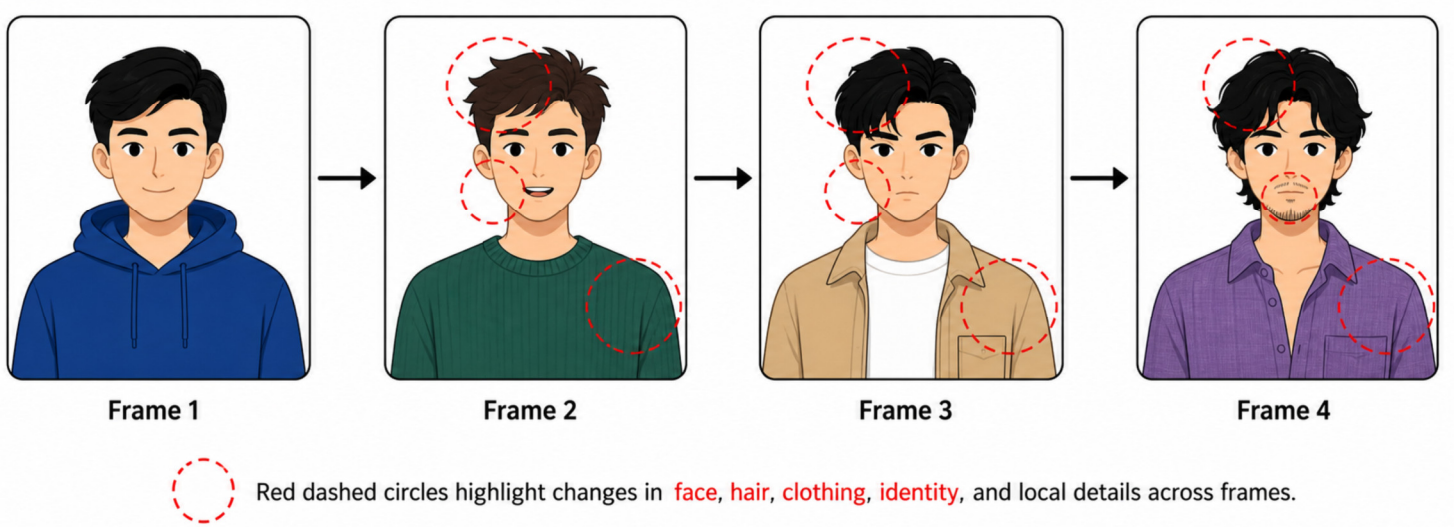


角色一致性的难点

跨帧中的脸部、服装、身份和局部细节容易发生漂移

漂移表现：

- 前后不像同一个角色
- 服装颜色/样式变化
- 面部特征不一致



解决方案方向：

- 参考图像保持（Reference Image）
- 身份嵌入（Identity Embedding）
- 跨帧注意力机制

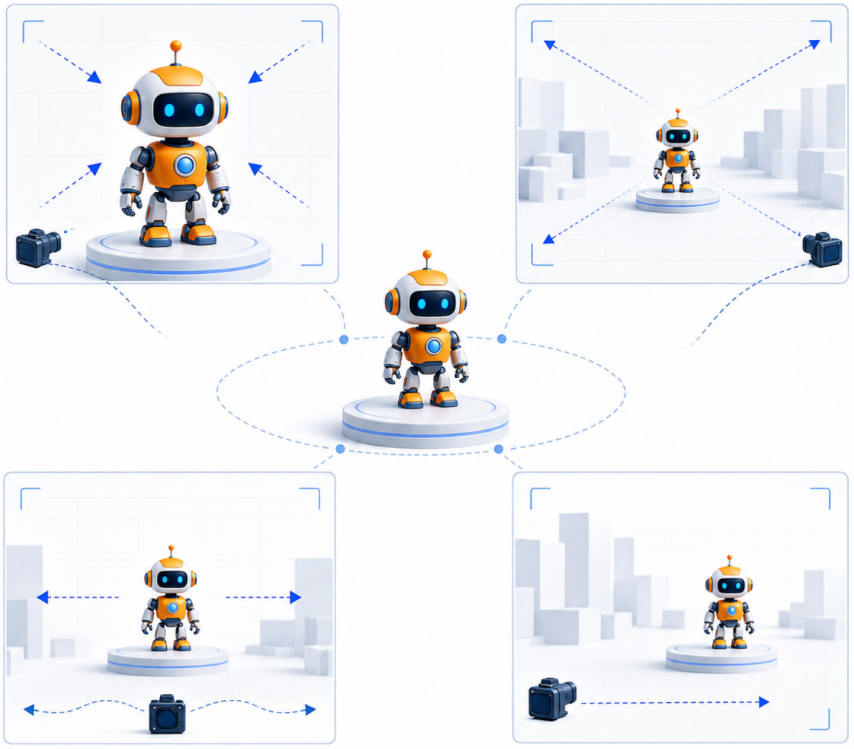
镜头控制与运动表达

高质量视频生成不仅要生成画面，

还要支持推拉摇移等镜头语言

镜头运动类型：

- **Push In / Pull Out（推/拉）**
- **Pan Left / Right（摇）**
- **Tilt Up / Down（俯仰）**
- **Dolly / Truck（移）**



视频生成 = "生成镜头叙事"

镜头语言是影视创作的核心表达手段，AIGC视频系统需要理解并支持这些运动模式

语音与音频生成

AIGC也能生成语音、配音、音乐和情感表达，形成完整内容链条

音频生成类型：

- 文本转语音（TTS）
- 音乐生成
- 音效合成

代表系统：Suno（音乐）、ElevenLabs（语音）、AudioLDM（音效）

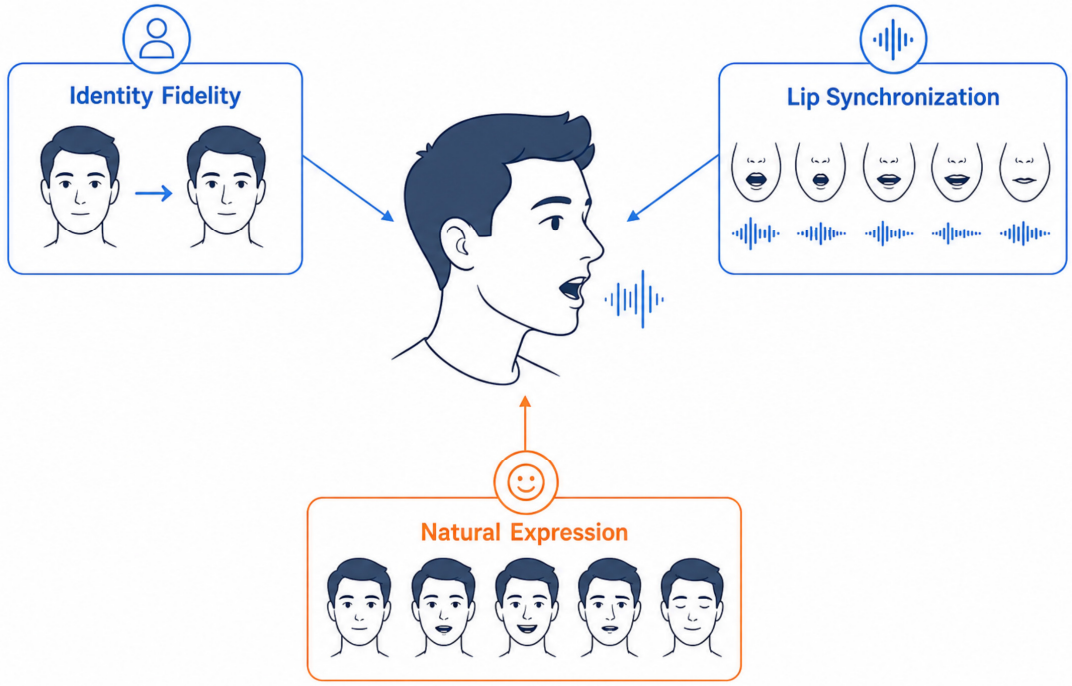
音频生成与图像/视频生成共享类似的扩散模型原理，但在信号表示和评估方法上有独特挑战

说话脸与数字人

说话脸生成要同时保证身份保真、口型同步和表情自然

技术挑战：

- **口型与语音精确对齐**
- **面部表情自然流畅**
- **头部姿态合理变化**



应用场景：虚拟主播、数字客服、在线教育、影视配音

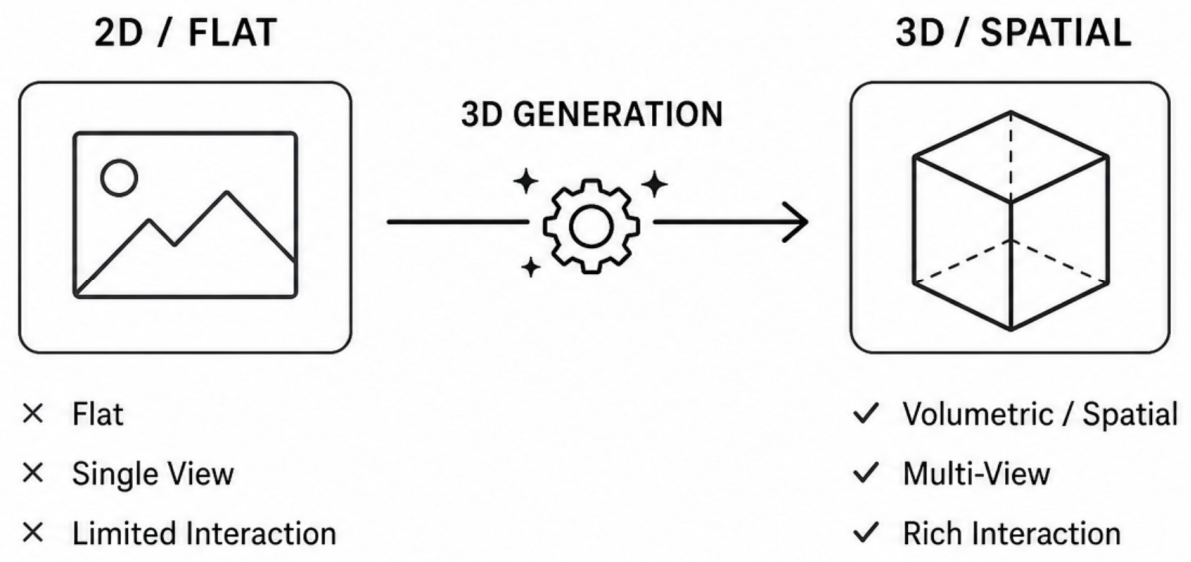
数字人是AIGC多模态能力的综合体现，需要图像生成、音频生成和时序建模的协同

3D生成为什么重要

3D生成使内容从"可看"走向"可交互、可导航、可仿真"

应用领域：

- 游戏开发
- VR/AR 体验
- 机器人仿真
- 工业设计



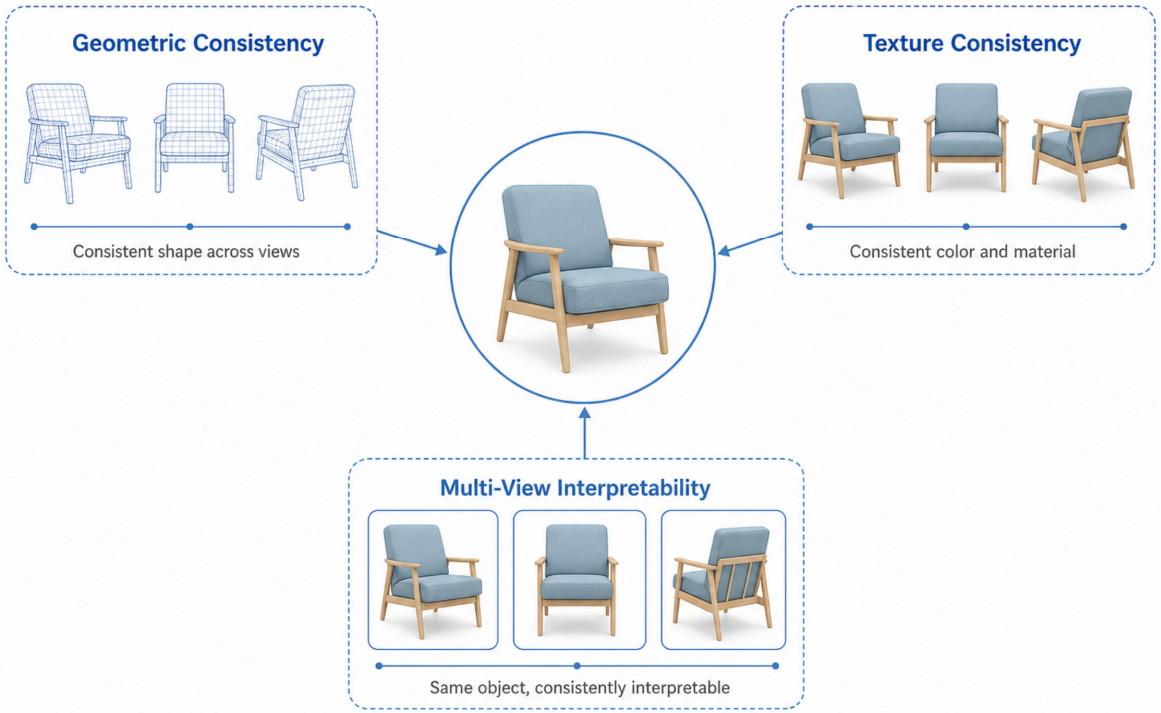
3D是AIGC的下一个前沿 — 从2D内容生成到3D世界构建，是实现沉浸式体验的关键技术

NeRF与文生3D的基本思路

3D生成需要同时解决几何一致性、纹理一致性和多视图可解释性

NeRF核心思想：

- **用神经网络表示3D场景**
- **输入：空间坐标+视角**
- **输出：颜色和密度**



文生3D路线：Text → 多视图生成 → 3D重建（NeRF/Gaussian Splatting）

代表工作：DreamFusion、Magic3D、Point-E

本节内容

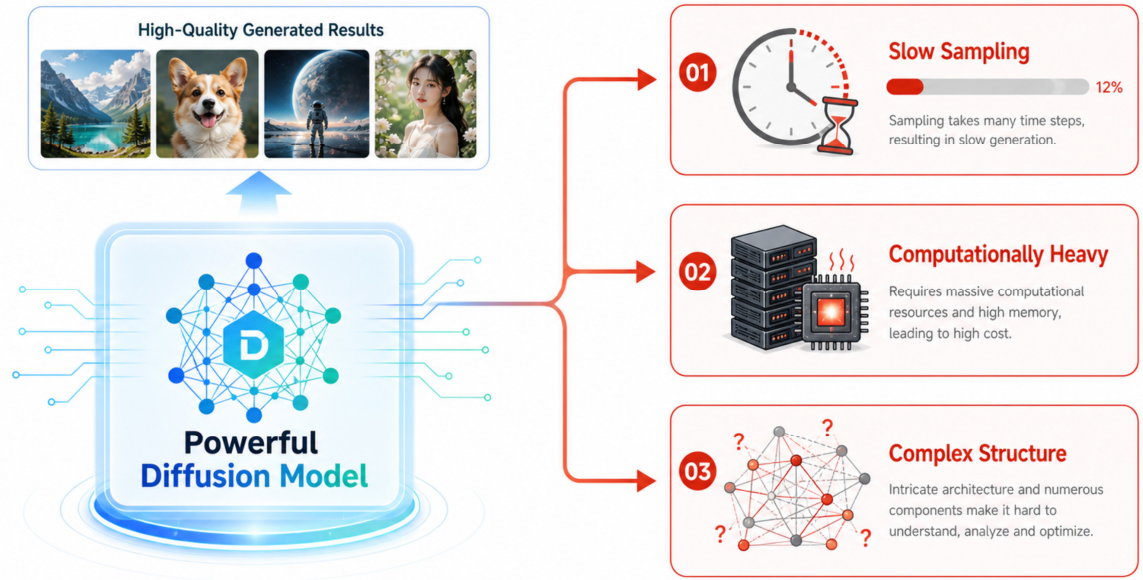
CONTENTS

为什么还要继续改进扩散模型

扩散模型虽然强大，但仍存在采样慢、计算重、结构复杂等问题

待解决问题：

- **推理速度：需要多步迭代**
- **计算成本：GPU资源消耗大**
- **架构简化：U-Net结构复杂**



改进方向：Flow Matching（统一视角）、Consistency Models（一步生成）、DiT（架构升级）、蒸馏加速

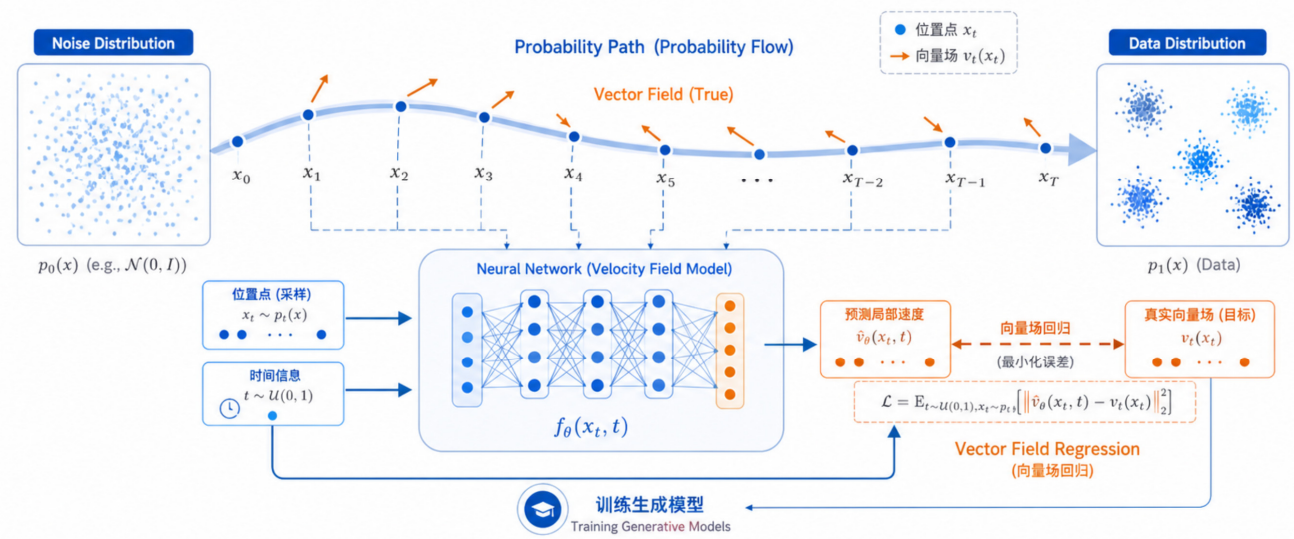
这些方向共同推动生成模型向更快、更简洁、更强大的方向发展

Flow Matching的基本思想

Flow Matching通过回归概率路径上的向量场来训练生成模型

核心区别：

- **扩散：模拟随机过程**
- **Flow Matching：回归向量场**
- **更简洁的数学框架**



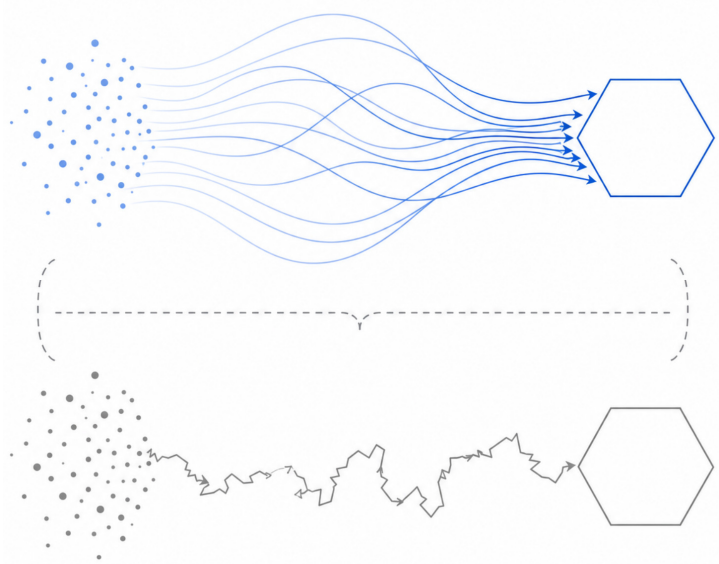
直观理解：不再模拟"噪声如何逐步加入"，而是学习"从数据到噪声的最优路径方向" — 就像学习一条河流的流速场

Flow Matching与扩散模型的关系

Flow Matching可以把扩散路径看作其特例，提供更统一的连续生成视角

扩散模型
- 定义前向随机过程
- 学习逆转该过程
- 特定的噪声调度

Flow Matching
- 定义概率路径
- 回归向量场
- 更一般化的框架



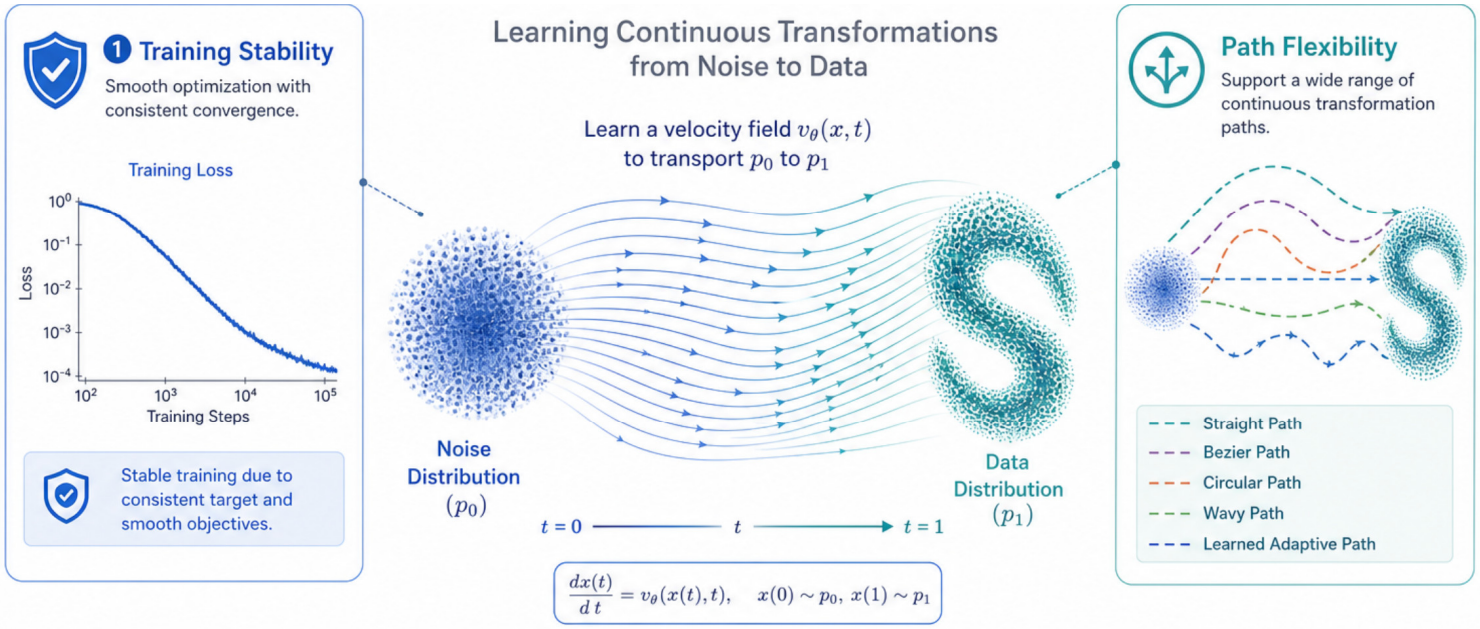
Flow Matching的优势

Flow Matching在训练稳定性、路径设计灵活性和高效采样方面具有吸引力

训练稳定
回归目标比去噪更稳定

路径灵活
可设计任意概率路径

高效采样
支持更少的采样步数

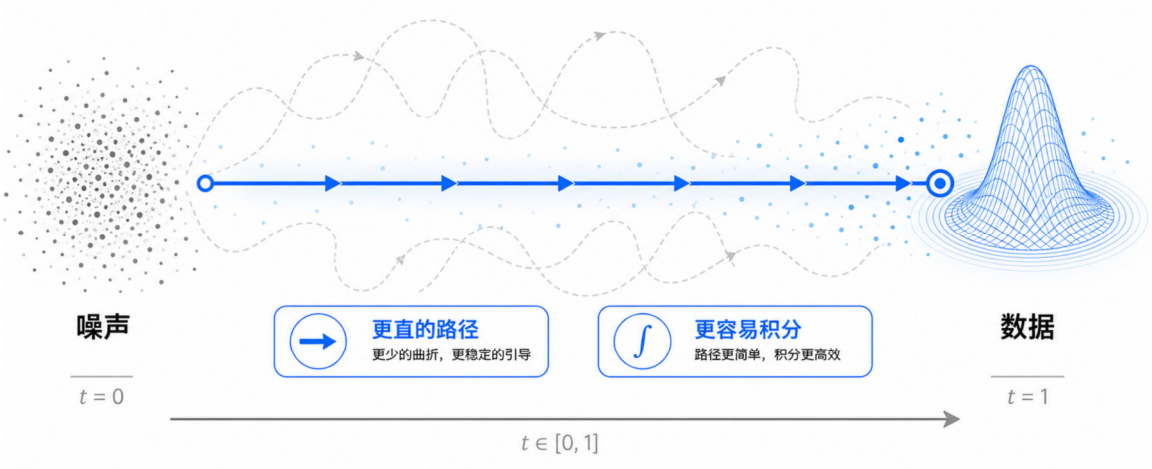


Rectified Flow的直觉

Rectified Flow强调让样本沿更直、更容易积分的路径从噪声走向数据

核心比喻：

- "把弯路拉直"
- 直线比曲线更容易走
- 减少采样步数



直观理解：想象从A点到B点，扩散模型走了一条蜿蜒的曲线（需要1000步），Rectified Flow把这条曲线拉直（只需几步就能到达）

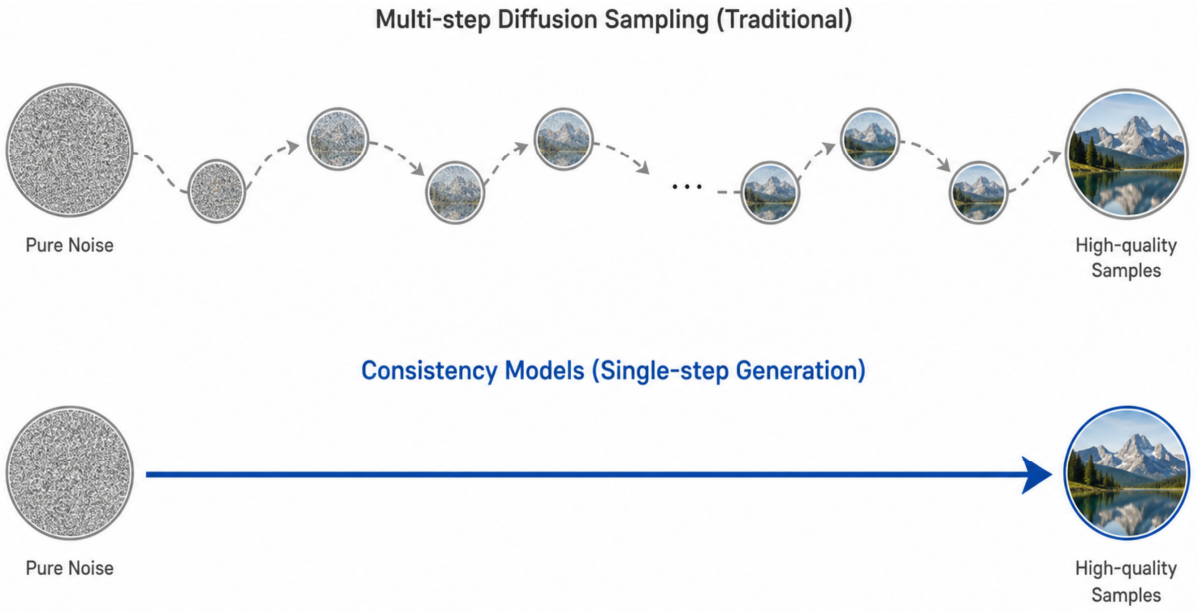
Consistency Models的动机

Consistency Models试图摆脱多步采样，

用一步或少步生成高质量样本

核心目标：

- 从"强质量"走向"快生成"
- 一步生成 vs 1000步生成
- 保持接近扩散模型的质量



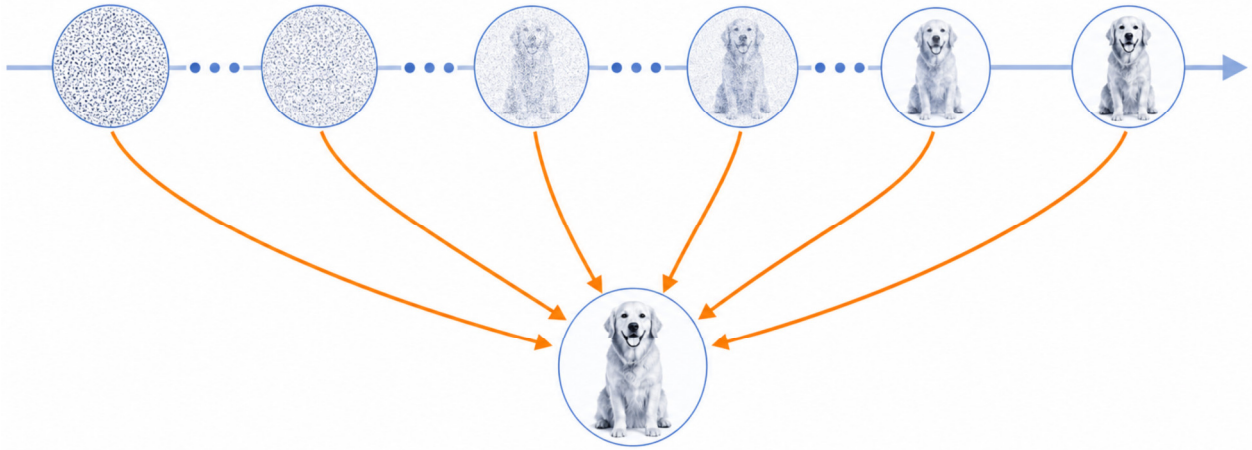
意义：如果成功，将彻底改变扩散模型的应用模式 — 从"等待数秒"到"即时生成"

Consistency Models怎么做

Consistency Models学习不同噪声层级之间的一致映射，支持快速采样

核心机制：

- 定义一致性函数 $f(x_t, t)$
- 任意噪声层级都映射到同一终点
- 训练时强制一致性约束



为什么少步采样成为可能：传统扩散需要逐步去噪，Consistency Models学会"跳步" — 直接从任意噪声状态预测最终干净样本

蒸馏与加速生成

很多快速生成方法的核心思想是把多步扩散过程蒸馏为少步甚至一步模型

Progressive Distillation

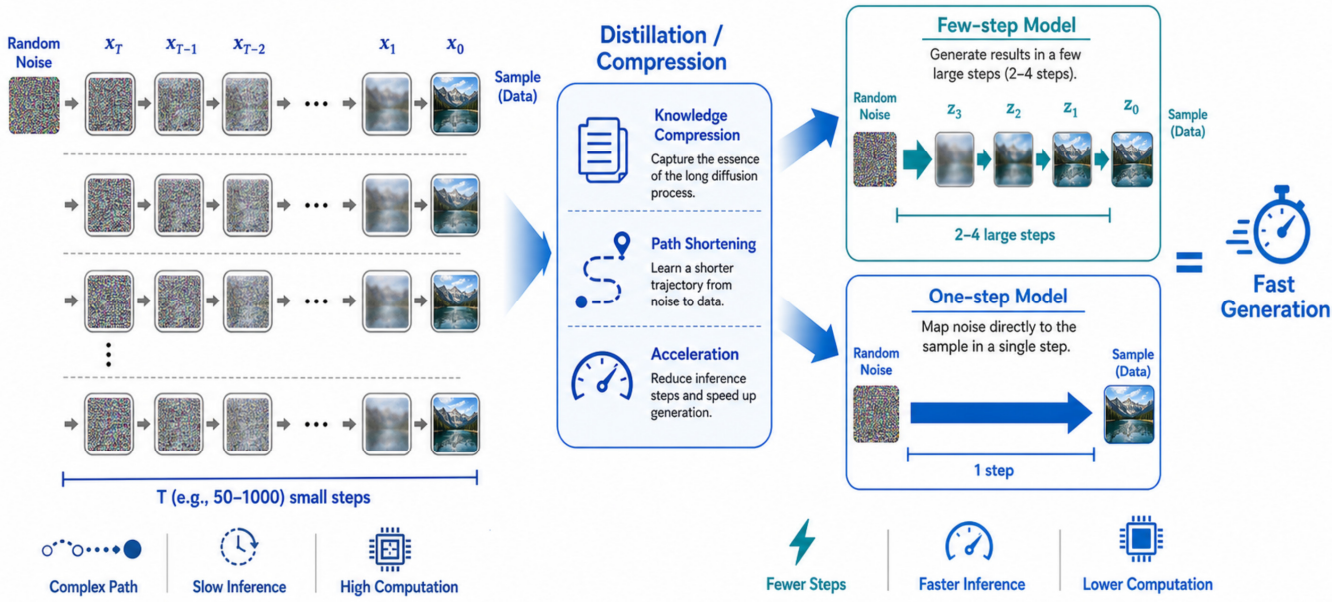
逐步将多步模型蒸馏为步数减半的学生模型

Guided Distillation

在蒸馏过程中保留CFG的控制能力

Consistency Distillation

结合Consistency Models思想的蒸馏方法

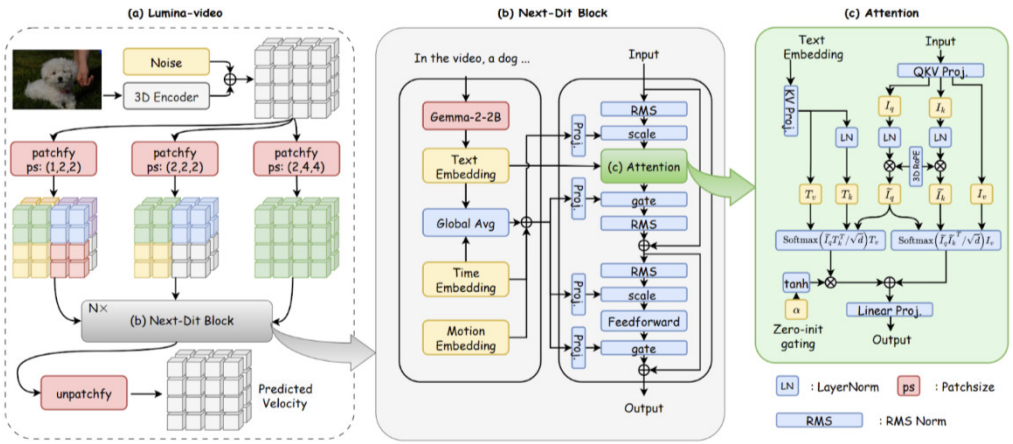


Diffusion Transformer（DiT）

DiT用Transformer替代传统U-Net作为扩散主干，体现生成模型的Transformer化趋势

核心改变：

- 将图像patch化为token序列
- 用Transformer处理
- 更好的扩展性



Architecture of Lumina-Video with Multi-scale Next-DiT and Motion Conditioning.

意义：DiT证明了Transformer在生成任务中的优势，为Sora等大规模视频生成系统奠定了架构基础

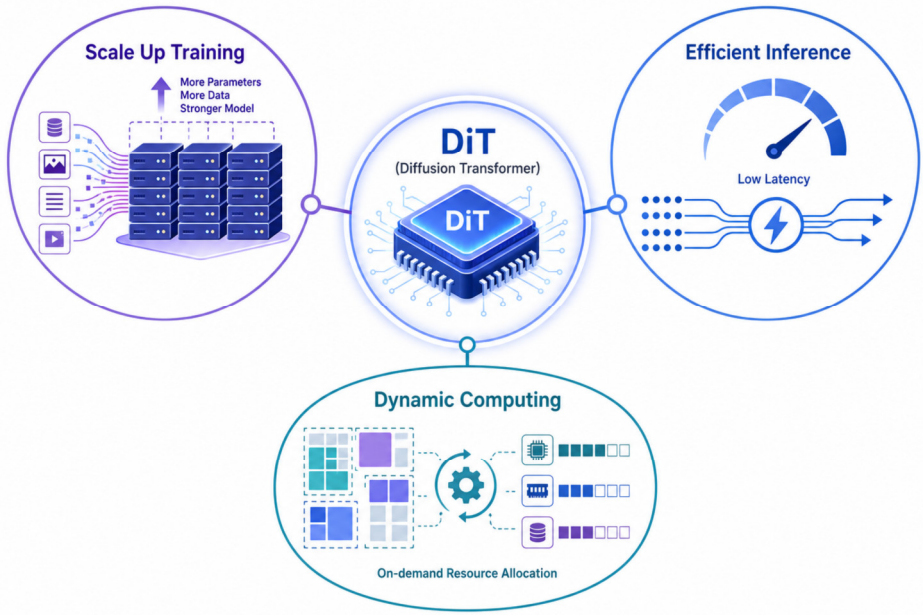
DiT的扩展方向

围绕DiT的研究重点包括更大规模训练、更高效率推理和动态计算

规模扩展
更大参数、更多数据

效率优化
高效注意力、量化压缩

动态计算
自适应深度、条件计算

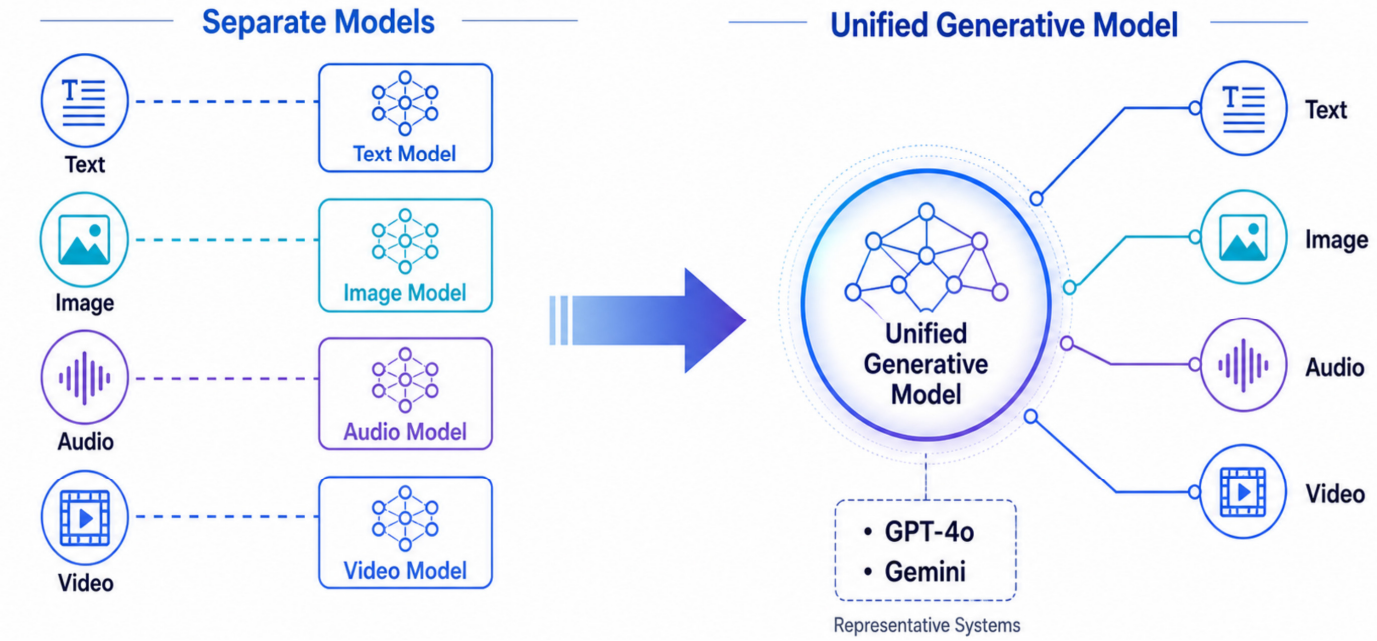


统一多模态生成的趋势

越来越多系统开始统一处理文本、图像、语音和视频

趋势特征：

- 单一模型处理多种模态
- 模态间无缝转换
- 统一表示空间



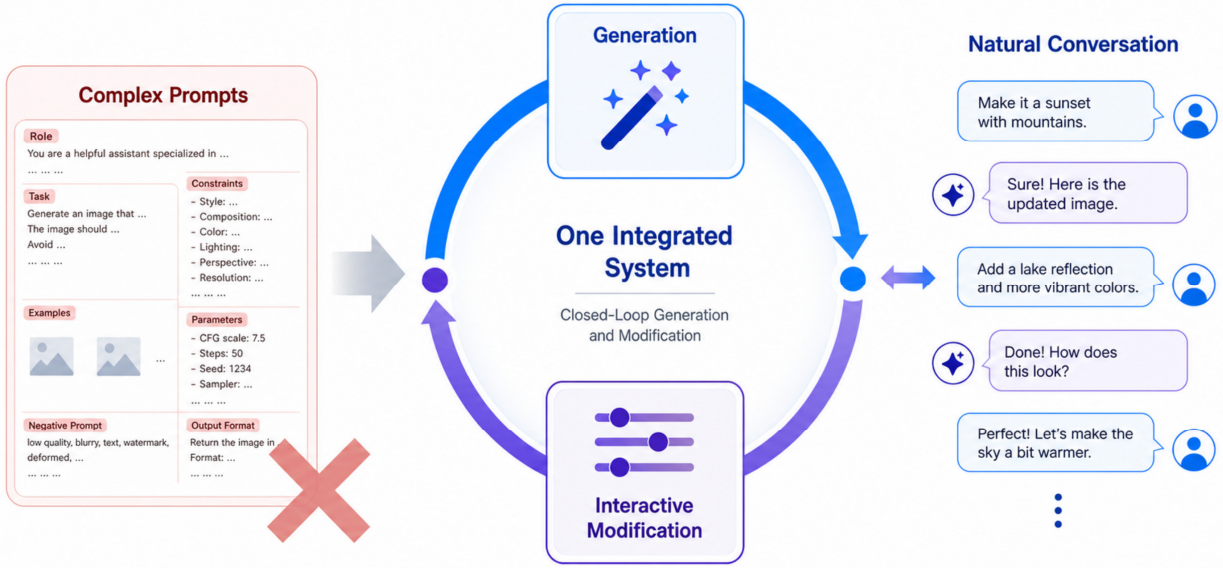
从"每种模态单独建模"到"统一生成模型" — 这是AIGC发展的重要方向，GPT-4o、Gemini等系统正在推动这一趋势

GPT-4o原生图像生成案例

新一代多模态系统开始把图像生成和对话理解统一到同一交互框架中

GPT-4o的特点：

- **对话式图像生成**
- **迭代修改和精调**
- **文本和图像统一理解**



"生成"与"交互式修改"走向一体化 — 用户不再需要写复杂的提示词，而是通过自然对话来指导生成过程

Sora案例：视频生成的新阶段

现代视频生成系统从"会动起来"走向"更可控、更连贯、更像真实镜头"

Sora的突破：

- 长视频生成（60秒）
- 复杂场景理解
- 物理规律遵循

技术基础：大规模DiT + 视频tokenization + 海量视频数据训练

Sora代表了视频生成从"研究模型"到"生产工具"的跨越

Veo案例：平台化视频生成

视频生成从研究模型走向开发平台和内容生产工作流

平台化特征：

- **API服务化**
- **与创作工具集成**
- **企业级部署**

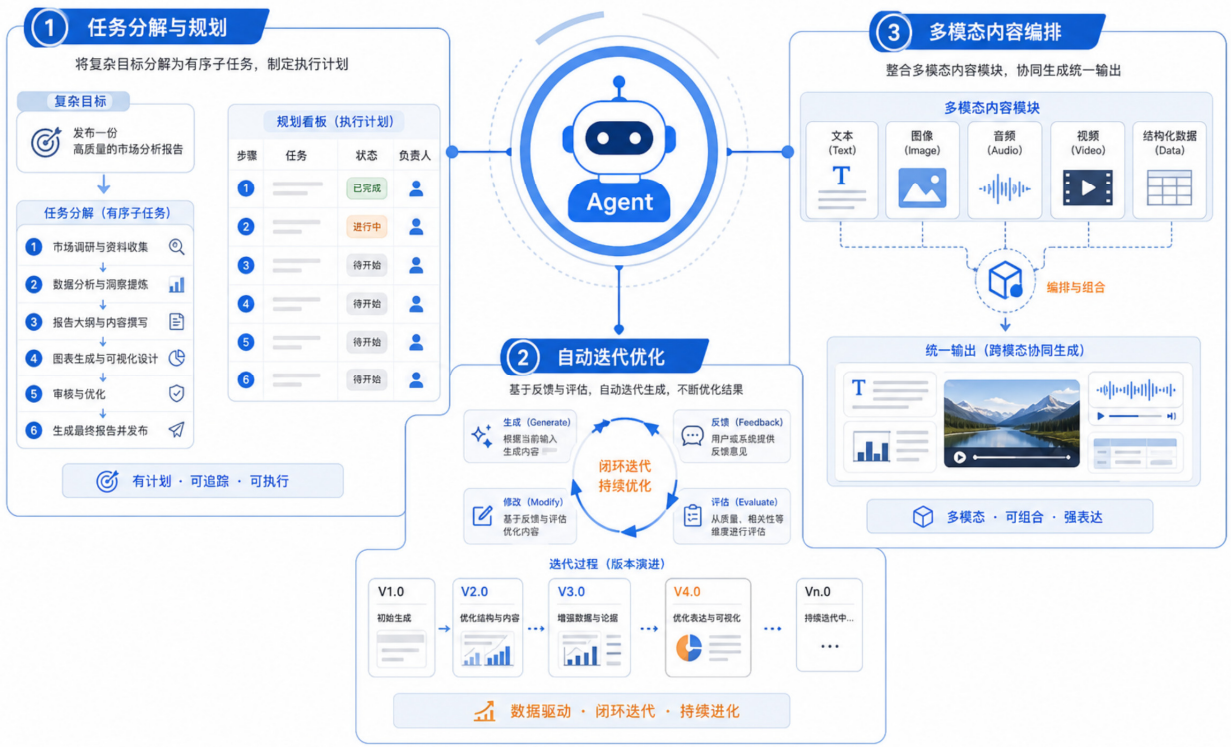
"系统能力"和"产品能力"已成为重要竞争点 — 模型本身不再是唯一壁垒，工程化、产品化和生态建设同样关键

AIGC与Agent工作流

未来的生成系统不再只是单次出图，而是能拆解任务、自动迭代和组织多模态输出

Agent + 生成：

- **任务分解与规划**
- **自动迭代优化**
- **多模态内容编排**



"生成 + Agent"可能形成新型内容生产链 — 从"人操作工具"到"Agent自动执行"，内容生产效率将发生质变

从内容生成走向世界生成

AIGC的长期趋势是从生成单个对象、单张图像，扩展到生成完整环境和交互世界

演进路线：

- **图像 → 视频 → 3D**
- **单对象 → 完整场景**
- **静态 → 可交互世界**

世界模型（World Models）是AIGC的终极愿景 — 生成的不只是内容，而是完整的可交互虚拟世界

问题和讨论