1. 流形假设

- 高维空间的图像数据，不会零散的分布在整个高纬空间，而是依附在一个低维的流形上。这个流形是数据的专属区域，只包含数据的核心信息，比如物体的形状，颜色，剔除了高维的冗余和噪声。
- 图像数据看起来是高维的，但本质上是低维的，像高维空间卷起来的一张低维纸，所有数据都落在这张纸上。

2. 传统扩散模型的弊端

- epsilon-pred，预测epsilon的方法（sd类），在高维下会失效，在预测整个高纬空间中的数据分布时，模型需要记住每个patch的特征，需要广大的网络宽度与冗余参数；
- v-pred，预测数据噪声混合流速，但带噪量v需要横跨整个高维空间，模型需要学习的范式更加复杂，也需要高容量，训练难度也高；

3. JiT 贡献

- JiT直接预测干净数据（x-prediction）
- 无需模型记高维噪声细节；
- 从带噪输入筛选低维干净数据特征；
- 降低模型容量需求，避免高维空间维度灾难；
- 契合精准提取核心信息逻辑；

4. JiT - 核心预测目标 （x- prediction）

- 输入带噪数据 z_t (z_t = t·x + （1-t)·epsilon), 模型直接输出干净数据预测值x_theta（x_theta = net_theta(z_t, t)）， 无需反推；
- 依流型假设，模型筛流形核心特征；
- 隐藏层低于 patch维仍精准预测；
- 用 v-loss优化，x_theta 转 v_theta保稳定；
- 不改变直接预测 x 逻辑，免复杂加权；

5. JiT - 极简架构设计（大 patch Transformer）

- 图像划非重叠大 patch，降序列长度；
- 256 x 256 用16x16 patch（长256)；
- 512 x 512 用32x32 patch（长256)；
- 核心组件仅线性嵌入，无额外模块；
- 无Tokenizer、预训练等，像素端到端训；
- 大 patch 保留局部语义完整性，减计算量，分辨率不影响长度；
- 1024x1024 与 256x256 计算量相当

6. JiT - 瓶颈嵌入；

- patch 嵌入层加“降维 + 升维”线性瓶颈；
- 高维patch 先降维d‘，再升维至隐藏层（将高维patch（如16x16x3=768）先降到低维的d‘(可低至16维)，再升到transformer的隐藏层的维度)；
- JiT用logit- normal采样t调噪声水平；
- 噪声调度最优 miu = -0.8 ，衡稳与质量；
- 瓶颈促进模型学习低纬特征，契合流型假设；
- 瓶颈 32 - 512 维度时，FID提升1.3
- 瓶颈低至 16 维，无性能退化；

7. 训练与推理

- 训练流程；
  - 核心步骤
    采样t -> 生成z_t = t·x + (1 - t) · epsilon -> 模型输出x_theta -> 计算v_theta = (x_theta - z_t)/(1 - t) -> v-loss (L = E_t,x,epsilon || v_theta - v ||^2)优化 -> 更新参数
  - 组件状态
    Transformer权重可训练，patch嵌入层（linear patch embedding）/位置编码（positional embedding）参数固定；
  - 关键操作
    含v-loss计算与梯度反向传播更新（更新Transformer核心权重与线性预测头权重）

- 推理流程
  - 核心步骤
    初始化z_0（噪声，z_0 ~ N(0, I) -> t从0到1迭代 —> 模型输出x_theta -> 计算v_theta=(x_theta - z_t)/(1-t) -> ODE求解 z_t+delta t -> t = 1时输出x_theta
  - 组件状态
    所有组件（Transformer、patch 嵌入层、位置编码、线性预测头）权重固定，仅执行前向计算
  - 关键操作
    含50步Heun ODE求解（d z_t / dt = v_theta(z_t, t)），保证生成平滑；