条件扩散与潜在空间生成

机器学习 · 生成模型专题

学习目标

- 区分带噪状态、时间、条件、预测输出和模型参数，解释条件训练与生成。
- 推导高斯扰动下得分与最优噪声回归的关系，以及条件引导与CFG的联系。
- 解释潜在表示、文本条件、U-Net / DiT和采样器如何组成生成系统，读懂原系统图并区分生成状态与插值结果。
- 推导所选DDIM构造与更新，说明时间子序列、随机性和边界。

得分与条件引导、DDIM构造分别独立展开。时间不足时可保留结论接口，完整推导材料继续保留。

可控、高效生成系统的四个问题

系统问题	机制	改变的对象
生成什么，条件作用多强?	条件训练与CFG	预测输入及采样使用的预测
在哪个空间重复计算?	潜在扩散LDM	像素状态改为潜在状态
什么网络负责预测?	U-Net或DiT	同一预测接口的内部结构
怎样减少采样调用?	DDIM	时间选择与状态更新规则

共同接口：当前状态、时间、条件进入预测器；预测交给采样器，产生较早状态。

四种选择可以组合。LDM可不使用CFG；DiT与DDIM分别承担网络和采样职责。

1 条件训练与CFG

生成什么，以及怎样调节条件作用?

固定条件与固定起点的两种对照

控制变量教学构造：固定同一模型、采样日程与确定性更新。

条件与起点	初始噪声A	初始噪声B
条件：圆形	 圆形，位置偏左	 圆形，位置偏右
条件：方形	 方形，位置偏左	 方形，位置偏右

横读一行：条件相同，改变初始随机性，仍可得到不同样本。

竖读一列：初始随机性相同，改变条件，观察指定属性如何变化。 本例只演示对照方法，位置对应是人工设置，不是扩散模型实验或“同种子必然保持布局”的保证。

论文真实生成结果



“毕加索风格的僵尸”
两幅输出；未控制随机来源

真实例来源：LDM v2，Fig.5，PDF p.6；200步、η=1，原文引导参数10。

条件提供指定信息，噪声提供随机变化

条件生成学习给定c后的数据分布$p(\boldsymbol{x}_0 \mid c)$。$c$可以是类别、文本或其他给定信息；它限制所讨论的样本范围，同一条件仍可对应多个结果。

继承离散加噪记号：$\alpha_t = 1 - \beta_t$、$\overline{\alpha}_t = \prod_{j=1}^{t} \alpha_j$，时间0为数据、T为噪声。为分开幅度与方差，记

$$a_t = \sqrt{\overline{\alpha}_t}, b_t = \sqrt{1 - \overline{\alpha}_t}, \boldsymbol{x}_t = a_t \boldsymbol{x}_0 + b_t \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}).$$

条件噪声预测网络$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c)$包含以下对象：

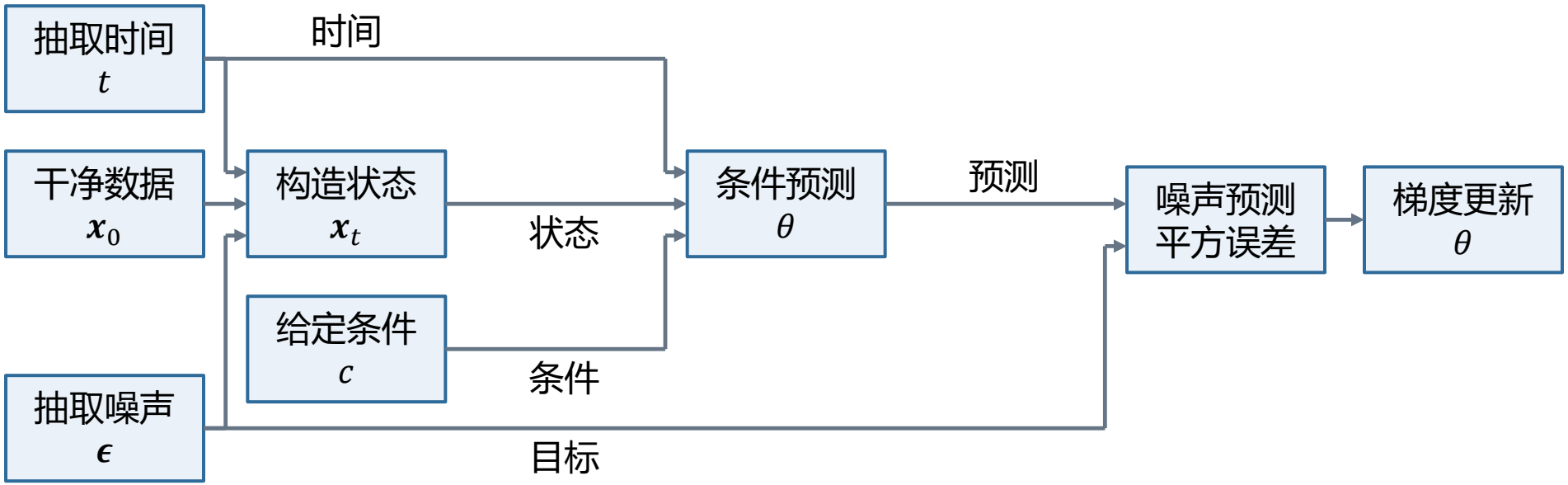
- 网络接收本次计算的带噪状态、时间和条件。
- 网络输出与状态同形状的噪声预测。
- 各次计算共享参数θ。

生成时给定c，干净数据$\boldsymbol{x}_0$尚未知。初始噪声和规定的采样随机源仍能产生不同结果；条件输入本身不保证结果准确符合条件。

配对数据形成条件训练项

训练取得配对数据$(\boldsymbol{x}_0, c)$，独立抽取时间与标准高斯噪声，构造已知目标：

$$L_{\text{cond}} = \mathbb{E}_{\boldsymbol{x}_0,c,t,\boldsymbol{\epsilon}}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(a_t\boldsymbol{x}_0 + b_t\boldsymbol{\epsilon}, t, c)\|^2.$$



训练用真实数据构造目标并更新参数；生成固定参数，网络据当前状态和条件预测噪声，采样器据此更新状态。

条件输入规定可用信息；若要在生成时调节其作用强度，还须修改预测。下一页先建立CFG的预测组合机制；其与得分的数学联系在深化A中推导。

CFG用预测差调节条件作用

已有条件预测后，怎样在生成时调节条件作用？固定同一$\boldsymbol{x}_t, t$，分别计算空条件预测$\boldsymbol{\epsilon}_{\mathrm{u}}$与真实条件预测$\boldsymbol{\epsilon}_{\mathrm{c}}$。 $\hat{\boldsymbol{\epsilon}}_s = \boldsymbol{\epsilon}_{\mathrm{u}} + s(\boldsymbol{\epsilon}_{\mathrm{c}} - \boldsymbol{\epsilon}_{\mathrm{u}})$.

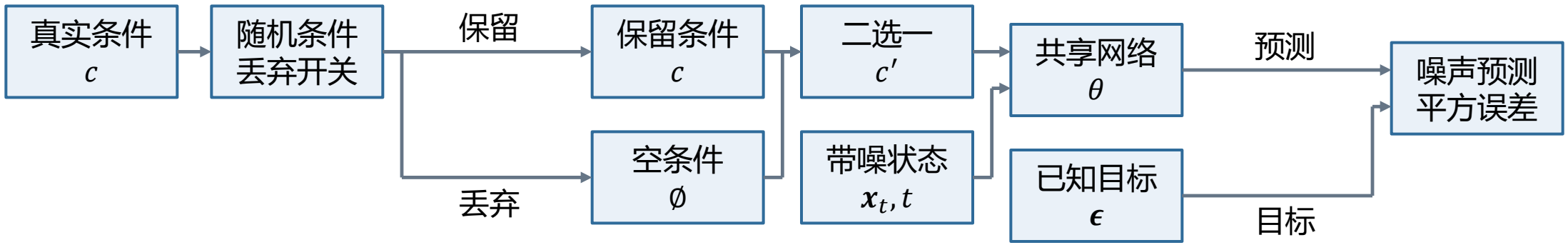
引导系数	使用的预测	几何含义
$s = 0$	$\boldsymbol{\epsilon}_{\mathrm{u}}$	无条件基点
$s = 1$	$\boldsymbol{\epsilon}_{\mathrm{c}}$	原条件预测
$s > 1$	沿两预测之差外推	加强这一预测差的作用

原CFG论文使用$(1 + w)\boldsymbol{\epsilon}_{\mathrm{c}} - w\boldsymbol{\epsilon}_{\mathrm{u}}$，因此$w = s - 1$。强引导不保证质量或多样性持续提高。理论依据完整保留在深化A；现在先解决两种预测怎样用一个网络获得。

条件丢弃让共享网络支持两种预测

设条件丢弃概率$p_{\text{drop}} \in (0,1)$。每次训练执行：

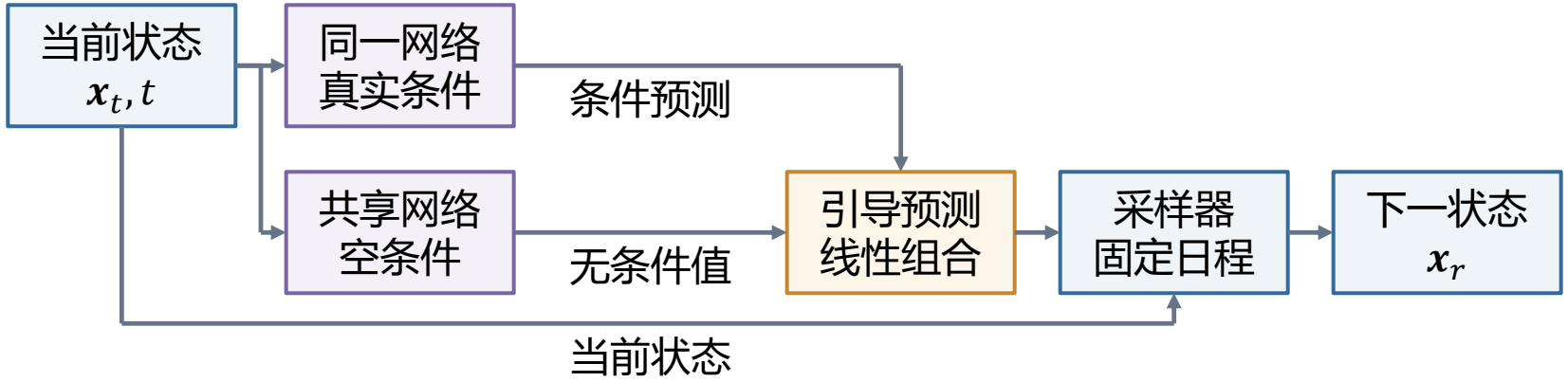
1. 抽取配对数据$(\boldsymbol{x}_0, c)$，以概率p_{drop}把条件换成空条件$\emptyset$，记结果为c'。
2. 抽取t与独立高斯噪声，构造$\boldsymbol{x}_t = a_t \boldsymbol{x}_0 + b_t \boldsymbol{\epsilon}$。
3. 计算$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c')$和噪声平方误差，更新共享参数$\theta$。



空条件分支学习对条件边缘化后的数据规律，保留条件的分支学习条件规律。两个分支共享模型；这里的“无分类器”仍包含条件数据训练。

采样时组合预测，再交给采样器

给定条件c和当前状态，固定训练后的参数，分别计算空条件与真实条件预测，再组合$\hat{\epsilon}_s$。随后把组合结果传给所选采样器；此步骤不更新参数。



取$\epsilon_{\mathrm{u}} = (0.2, -0.1)$、$\epsilon_{\mathrm{c}} = (0.5, 0.3)$，差为$(0.3, 0.4)$。$s = 0,1,2$分别得到$(0.2, -0.1)$、$(0.5, 0.3)$、$(0.8, 0.7)$。第三个结果是外推，不能据此保证更符合条件或更有多样性。

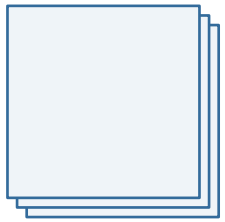
两个前向通常增加网络计算；实现可把它们合批，但样本、权重、步数与单次成本仍需共同记录。质量与成本的评价由具体设置决定。

2 潜在空间与条件预测网络

在较小状态空间重复预测，再把结果解码成图像。

重复预测的状态从像素改为潜在张量

教学尺寸假设：像素图像256×256×3，编码后潜在张量32×32×4，空间压缩因子为8。



256 × 256 × 3

每层：一个空间通道

像素空间：3个通道



32 × 32 × 4

单通道边长按 8:1 示意

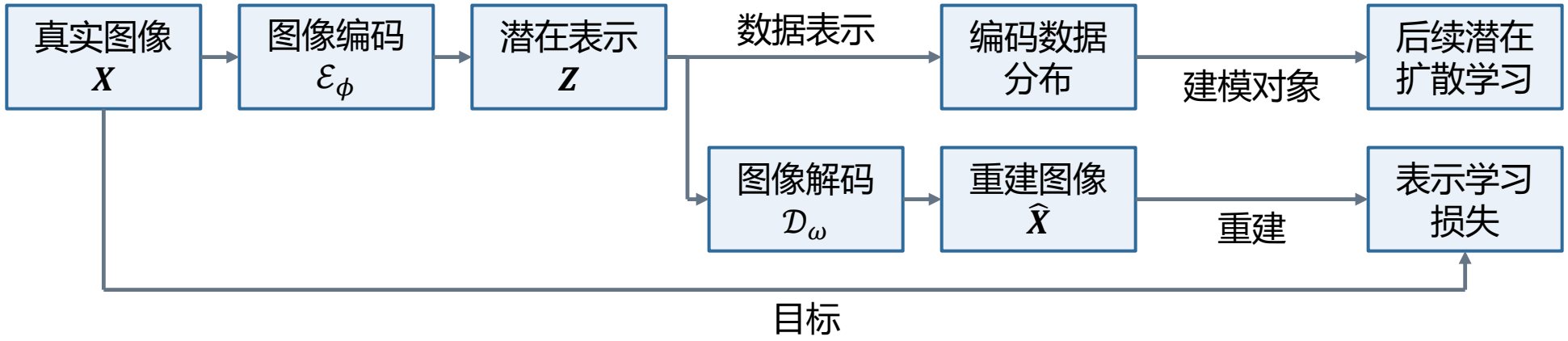
路径	重复去噪的张量	去噪结束后
像素扩散	256×256×3：196608元素	直接得到像素图像
潜在扩散	32×32×4：4096元素	恢复尺度，再调用图像解码器

元素数为原来的1/48，实际时间还取决于通道、骨干、注意力与硬件，不能据此声称加速48倍。

训练时真实图像先编码；从噪声生成时，重复循环只更新潜在状态，结束后解码一次。

表示学习与生成学习承担不同目标

像素空间的每一步预测都要处理高维图像。潜在扩散把重复预测移到较小的表示空间：先用表示模型压缩图像，再学习编码数据的分布。图像编码器$\mathcal{E}_\phi$把$\boldsymbol{X}$映射到潜在张量，解码器$\mathcal{D}_\omega$重建图像： $\boldsymbol{Z} = \mathcal{E}_\phi(\boldsymbol{X}), \widehat{\boldsymbol{X}} = \mathcal{D}_\omega(\boldsymbol{Z})$.

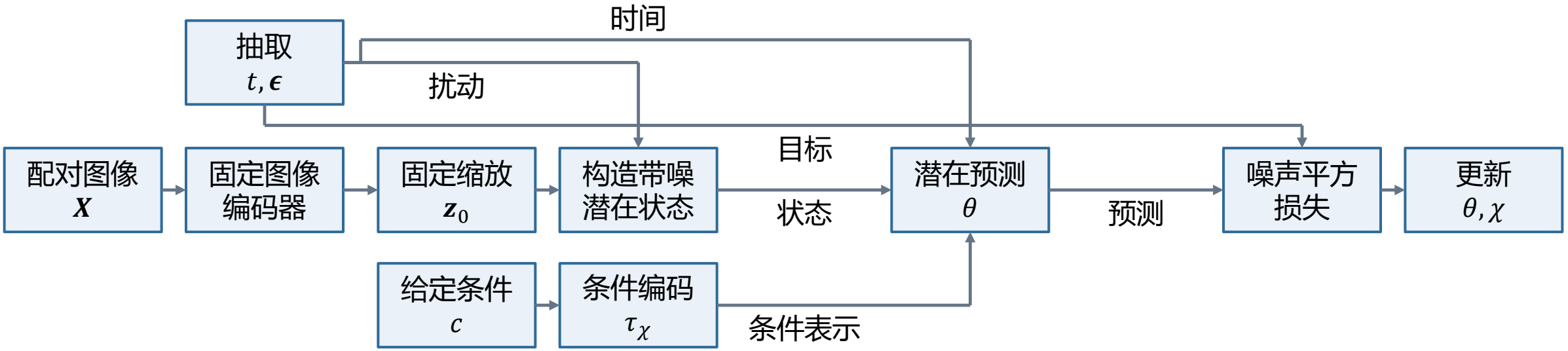


LDM原论文采用感知重建与局部对抗目标，并讨论轻度KL正则或VQ正则。具体目标决定表示保留什么信息；它不必等于GM02的高斯VAE目标。压缩存在信息损失，后续扩散也受表示与解码能力约束。

尺度一致后在潜在状态上加噪

取固定缩放$k > 0$，定义潜在扩散使用的干净状态$\boldsymbol{z}_0 = k\mathcal{E}_\phi(\boldsymbol{X})$。本讲基线先训练表示，再固定$\phi, \omega, k$，学习 $\boldsymbol{z}_t = a_t\boldsymbol{z}_0 + b_t\boldsymbol{\epsilon}, L_{\text{latent}} = \mathbb{E}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \tau_\chi(c))\|^2$.

这里噪声与潜在状态同形状，条件编码器τ_χ可按所选方案与去噪器联合学习；本讲采用这一明确方案。具体系统若冻结条件编码器，应相应去掉其参数更新。



固定缩放改变潜在信噪比

缩放改变信号相对噪声的幅度。若$Var(Z) = 4$，取$k = 0.5$后方差变为1。

在$a_t = 0.8, b_t = 0.6$时，信号方差与噪声方差之比从$0.64 \times 4/0.36$变为$0.64/0.36$。

固定k须训练和生成一致，本文不指定通用产品常数。

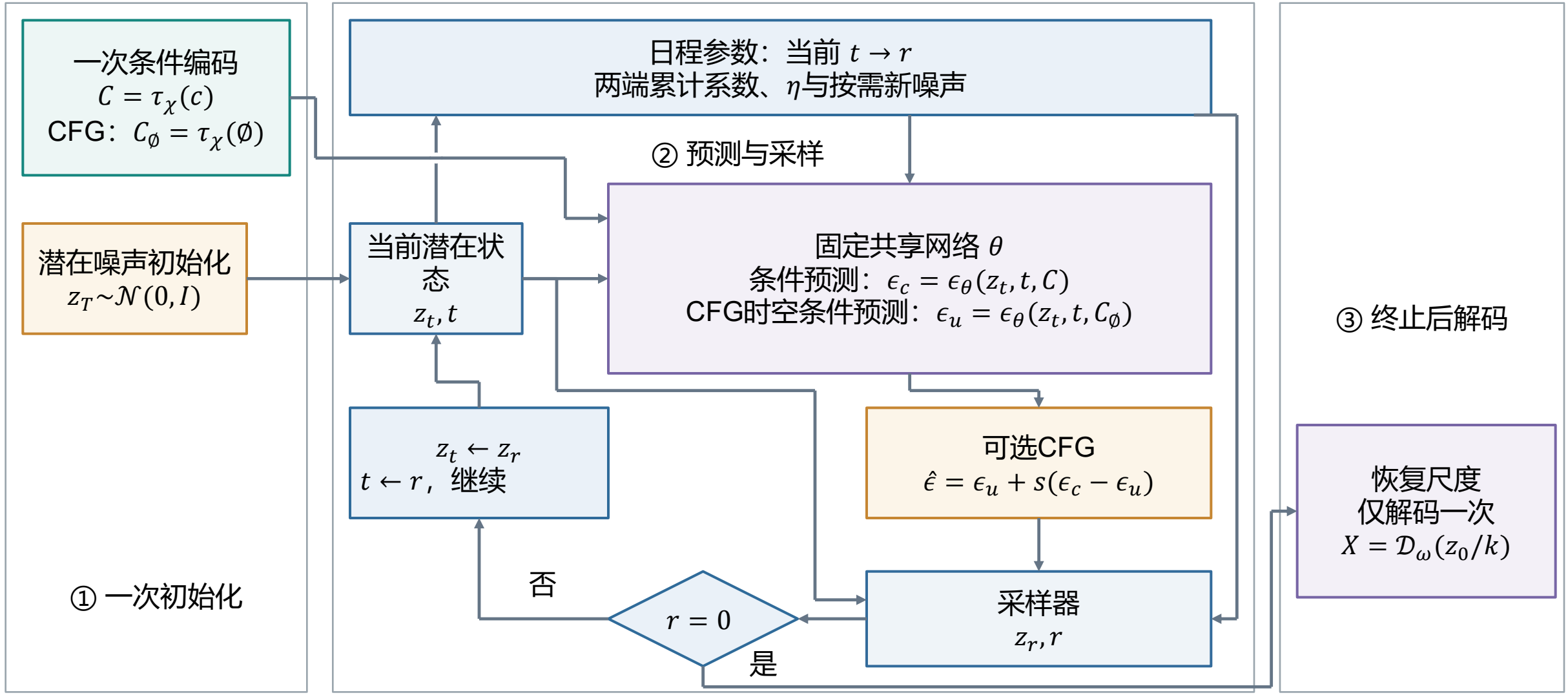
从潜在噪声生成，再恢复尺度解码

训练完成后，给定c并固定所有参数，执行：

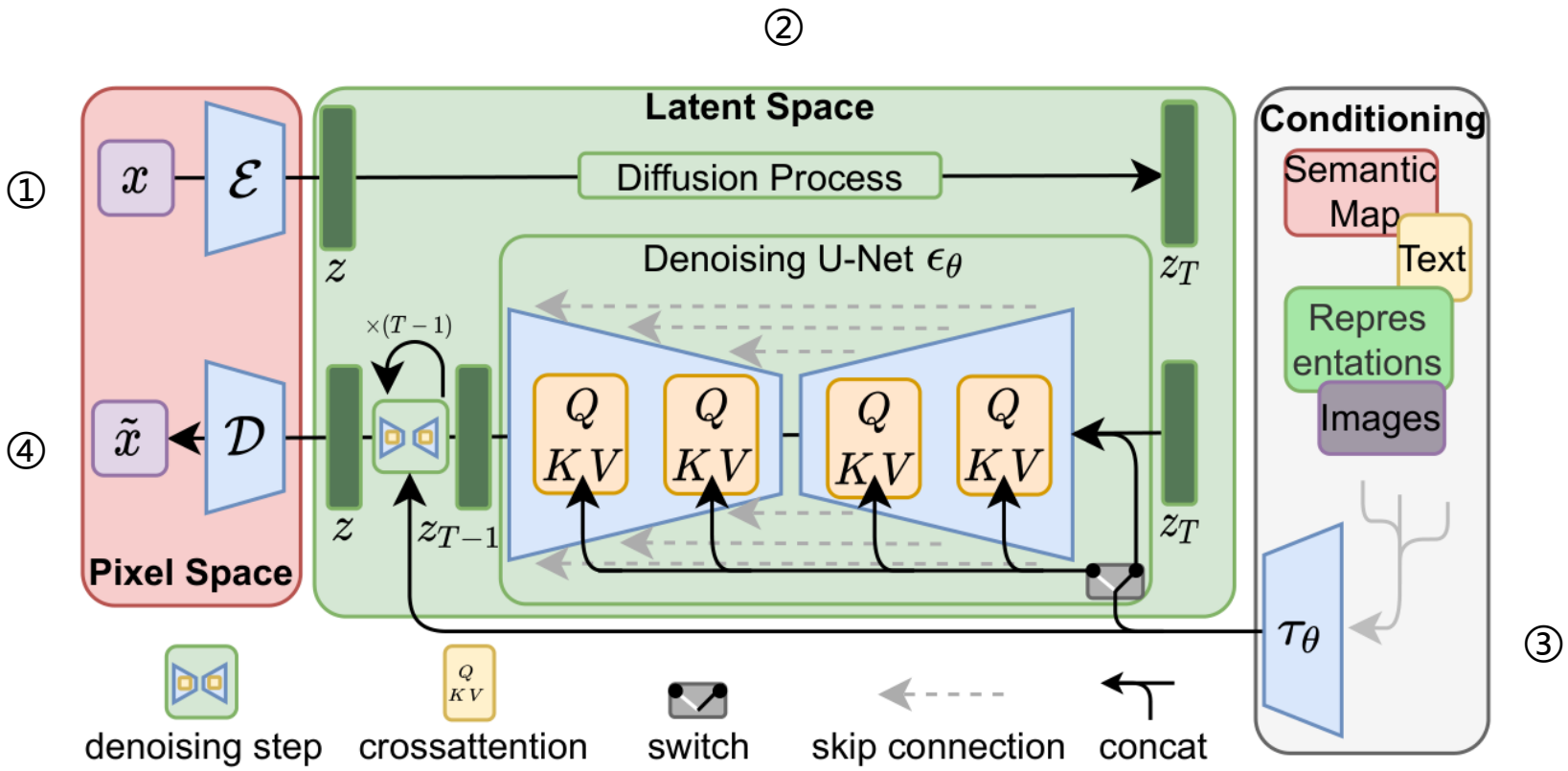
1. 计算条件表示$\tau_\chi(c)$，在潜在空间抽$\boldsymbol{z}_T \sim \mathcal{N}(0, \boldsymbol{I})$。
2. 按选定时间序列，使用$\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \tau_\chi(c))$，需要时组合空条件预测，再按采样器更新潜在状态。
3. 获得$\boldsymbol{z}_{\text{out}}$后，输出$\boldsymbol{X}_{\text{out}} = \mathcal{D}_\omega(\boldsymbol{z}_{\text{out}}/k)$。

从噪声生成的路径没有待重建图像，通常不调用图像编码器；图像到图像等其他任务可另有编码输入。本生成过程学习编码数据分布，承担的职责超出了从原AE简单先验直接解码。

潜在生成系统的组件与循环



原LDM图把潜在循环与条件支路放在一起



① 图像经 E 编码到潜在空间

② 潜在状态反复预测与采样更新

③ 条件编码提供交叉注意力 K/V

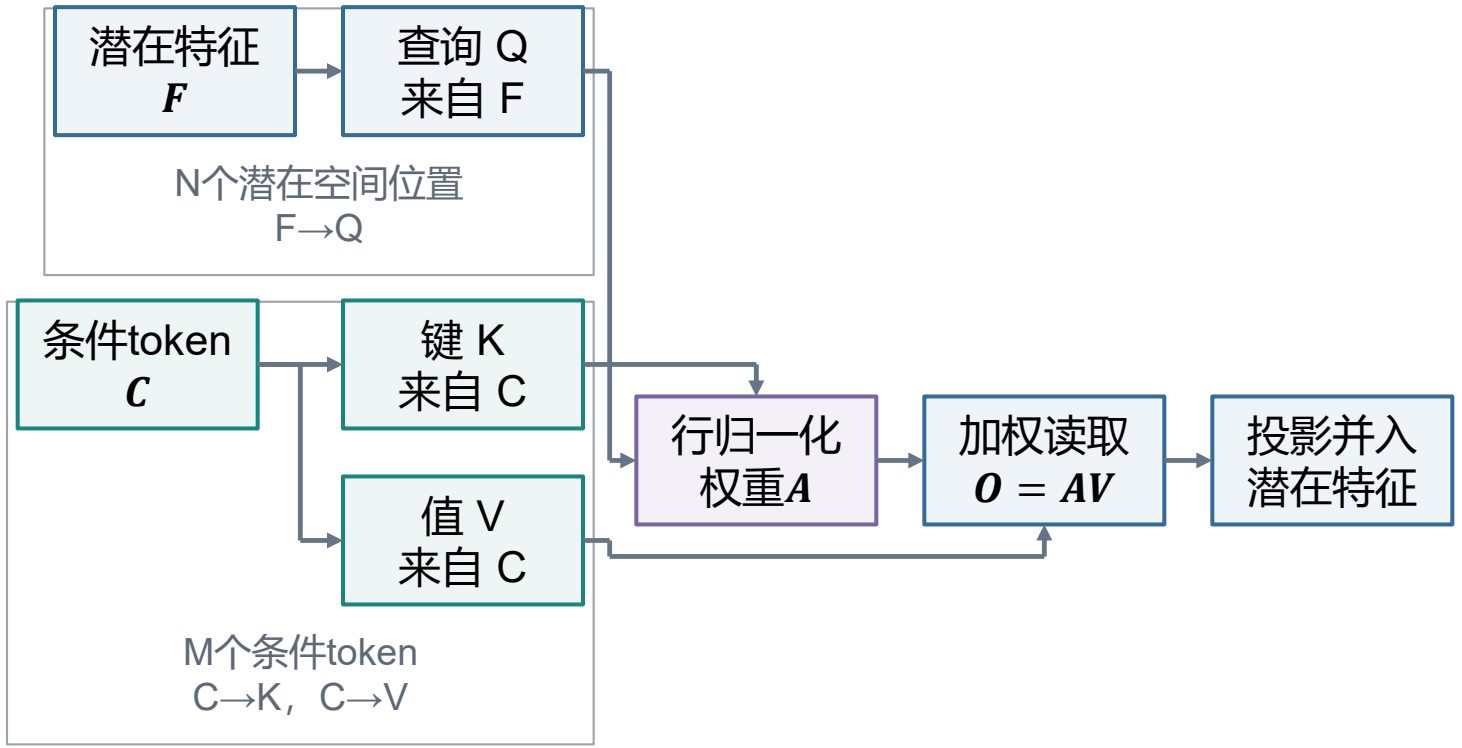
④ 循环结束后经 D 解码为图像

本课额外约定：训练前固定 E/D；生成结束先除 k，再解码。

来源：Rombach等，LDM，arXiv:2112.10752v2，Figure 3，PDF p.4；编号为教学辅助。

潜在特征提供查询，条件表示提供键和值

潜在特征$F \in \mathbb{R}^{N \times d_f}$按空间位置排列；条件$C \in \mathbb{R}^{M \times d_c}$按token排列。每处空间位置读取条件信息：$\boldsymbol{Q} = \boldsymbol{F}\boldsymbol{W}_Q, \boldsymbol{K} = \boldsymbol{C}\boldsymbol{W}_K, \boldsymbol{V} = \boldsymbol{C}\boldsymbol{W}_V, \boldsymbol{A} = softmax_{\text{row}}(\boldsymbol{Q}\boldsymbol{K}^{\top}/\sqrt{d_a}), \boldsymbol{O} = \boldsymbol{A}\boldsymbol{V}.$



各空间位置用自己的查询读取条件信息，因而条件能改变后续预测；关注权重不自动成为准确词物对齐或条件满足的保证。

一个空间位置怎样读取条件token

教学权重示例：2个空间位置查询3个文本token。权重由人工指定，只演示行归一化与加权聚合。



查询位置	红色	小猫	草地	行和
位置①	0.6	0.3	0.1	1
位置②	0.1	0.2	0.7	1

取二维值向量$V_1 = (1,0), V_2 = (0,1), V_3 = (1,1)$，则位置1读取 $O_1 = 0.6V_1 + 0.3V_2 + 0.1V_3 = (0.7,0.4)$. 位置2得到$O_2 = (0.8,0.9)$。两行结果分别返回各自空间位置，经输出投影并入潜在特征。

一般形状：$Q: N\times d_a$，$K: M\times d_a$，$V: M\times d_v$，$A: N\times M$，$O: N\times d_v$。权重高只表示此层读取较多，不保证词物准确对齐。

条件编码、时间编码与空间位置分别提供信息

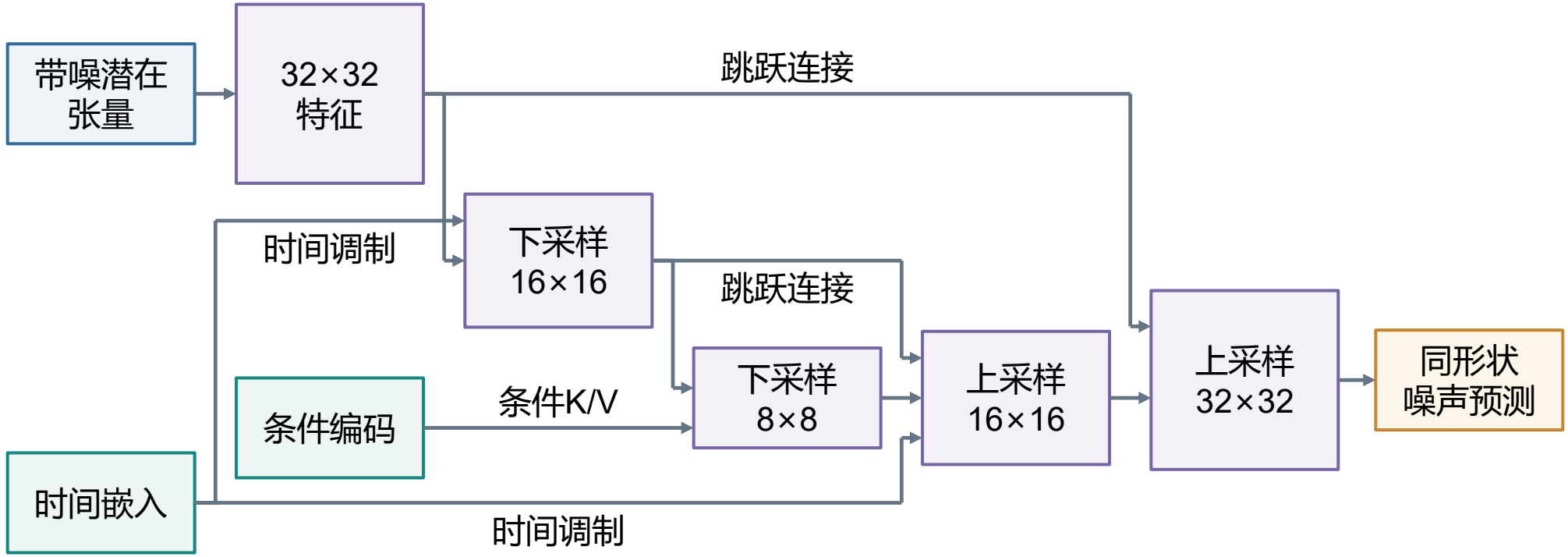
对象	提供的信息	与其他对象的区别
图像编码器$\mathcal{E}_{\phi}$	把图像映射到生成状态空间	从噪声生成时通常不接收待生成图像
条件编码器τ_{χ}	给定文本或标签的表示	不把给定文本当作潜在图像状态
时间编码	当前噪声等级	不表示图像中的空间位置
空间位置编码 / 空间结构	图像或patch所在位置	不决定当前噪声大小

在原LDM条件机制中，领域条件编码器可以与去噪网络联合优化。使用预训练并冻结的文本编码器是另一种系统决定，须按具体模型说明。

CLIP文本编码器是条件表示的一种来源；LDM框架不要求所有实现固定使用CLIP。拼接、交叉注意力和条件调制都是可选注入方式，需与所采用网络结构一致。

多尺度U-Net处理状态与条件

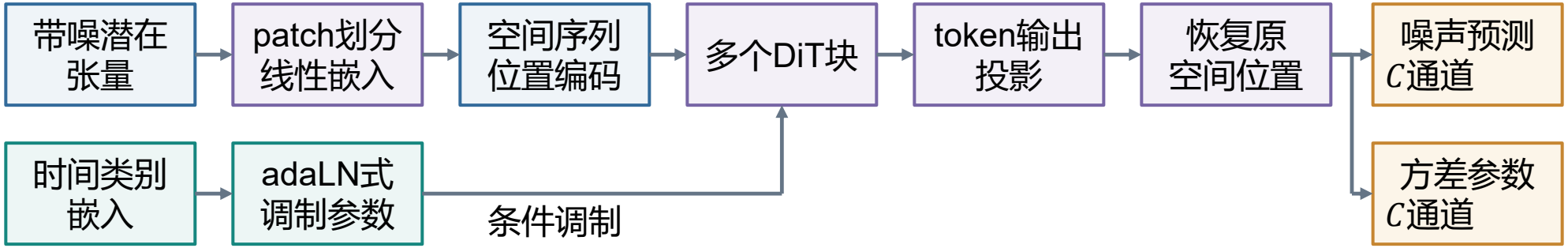
前面固定了状态、时间、条件和噪声输出的身份，现在看内部怎样实现这项预测。U-Net在多个空间尺度上处理带噪潜在张量，低分辨率层整合较大范围信息，上采样恢复空间分辨率，跳跃连接把相应尺度的特征送往后段。时间与条件输入调制计算。



图中32、16、8为教学空间尺寸，通道数与实际层数由模型决定。这张图只固定接口与跨尺度关系，不规定实际层数、通道数或每层都采用同一种条件注入。完整卷积、残差和注意力计算沿用DL先修。

DiT把潜在状态转成patch序列再还原

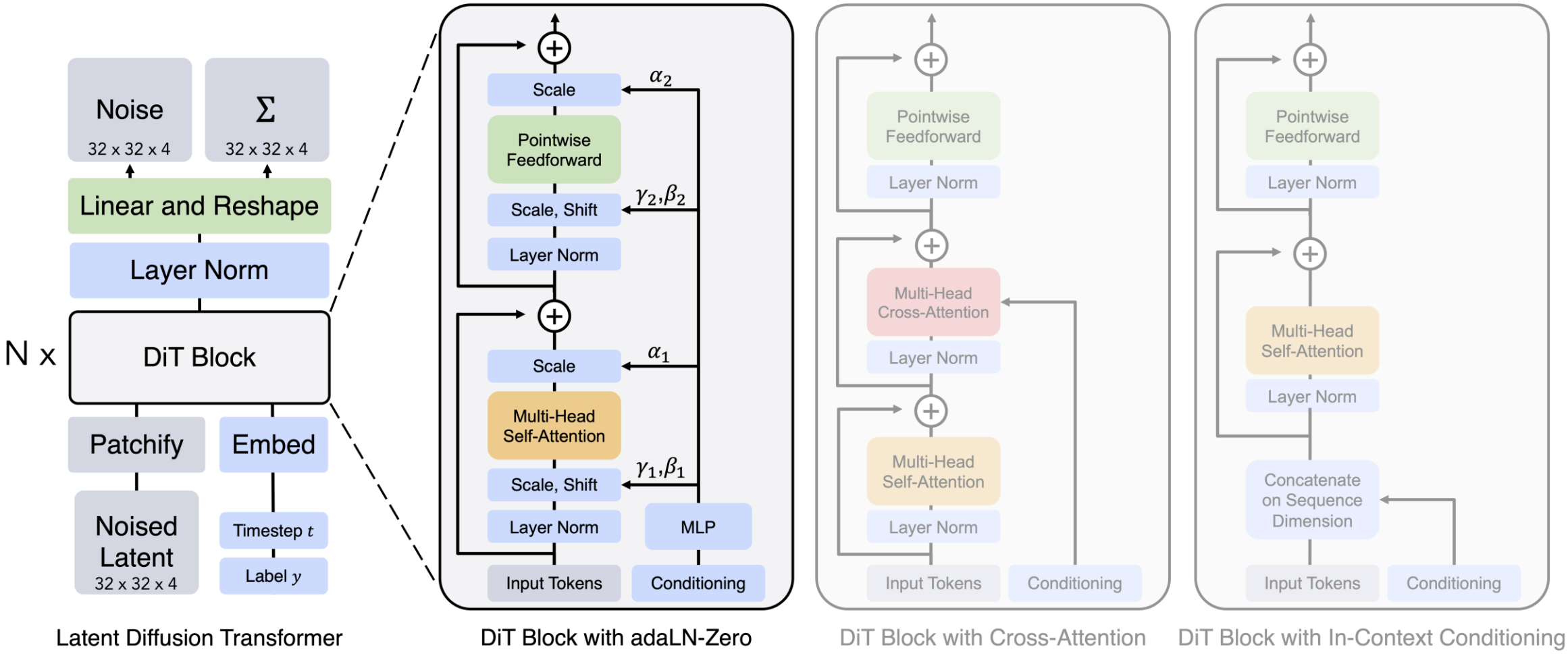
对$h\times w\times C$潜在状态，取能整除h、w的patch边长p，得到$N=(h/p)(w/p)$个patch。各patch线性嵌入后加入空间位置编码，送入Transformer块。



原DiT论文比较条件token、交叉注意力、adaLN和adaLN-Zero等设计。adaLN由时间与类别嵌入产生归一化后的缩放和平移，adaLN-Zero还控制残差分支，相关调制输出采用零初始化；并非把整个网络的所有权重置零。不同设计分别实现条件作用，所有DiT并非都依赖交叉注意力。

以$32\times32\times4$潜在状态、$p=2$为例，得到256个patch；若每个token输出$p^2(2C)=32$个量，重排后为$32\times32\times8$，分成两组4通道。原论文输出噪声及方差参数；本讲DDIM例只使用噪声支路并显式指定采样方差。

DiT原图中的条件注入设计



Latent Diffusion Transformer　DiT Block with adaLN-Zero　DiT Block with Cross-Attention　DiT Block with In-Context Conditioning

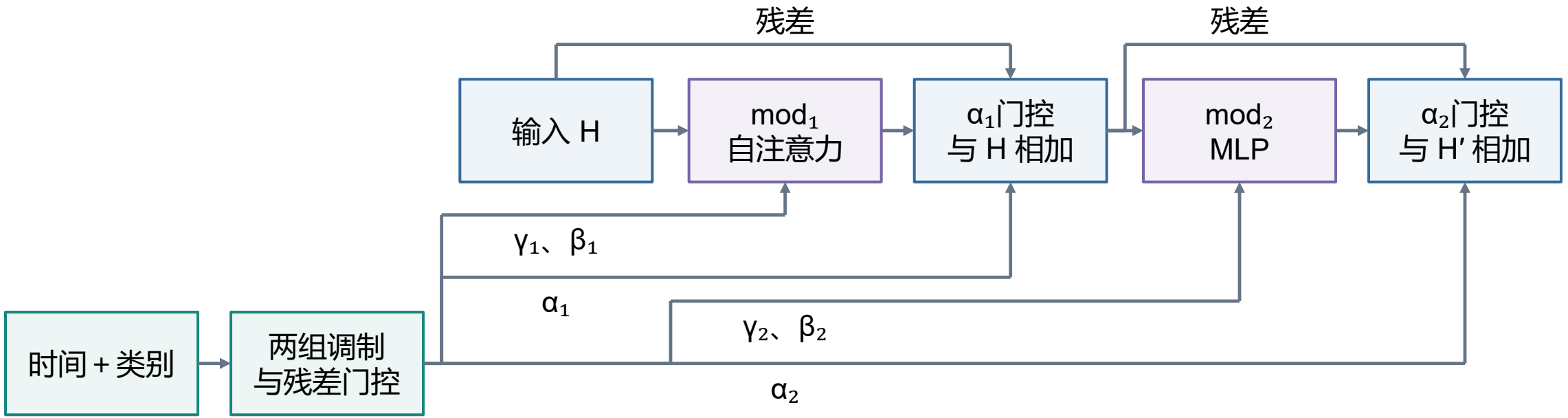
原始DiT：时间 t ＋ 类别 c；图中是不同条件注入候选，区别于P19的文本交叉注意力示例。

来源：Peebles与Xie，DiT §3.2/Figure 3，作者项目页 wpeebles.com/DiT；保留原配色。

adaLN-Zero把条件变成调制与残差门控

记联合嵌入$g = e_t + e_c$，来自时间t与类别c。两组调制分别作用于注意力与MLP：

$$mod_j(H, g) = (1 + \gamma_j(g)) \odot LN(H) + \beta_j(g),$$
$$H' = H + \alpha_1(g) \odot Attention(mod_1(H, g)),$$
$$H'' = H' + \alpha_2(g) \odot MLP(mod_2(H', g)).$$



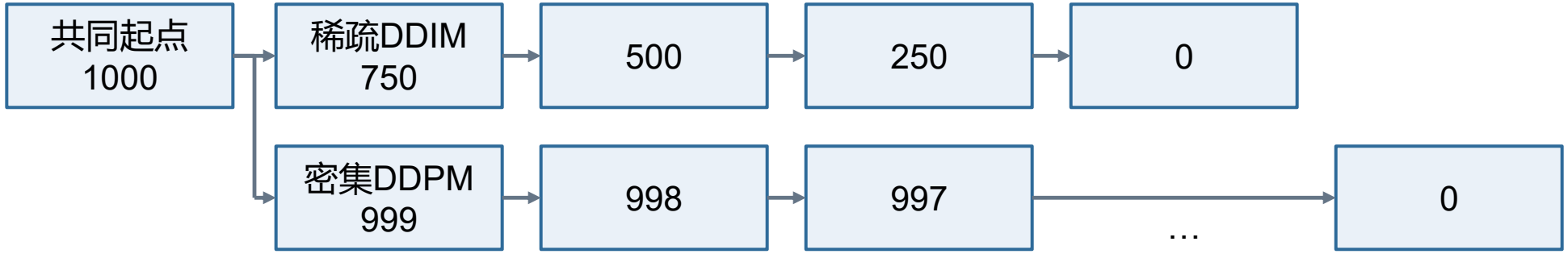
α_1=α_2=0使块初始为恒等映射；零初始化限于相关调制输出和门控，其余模型参数正常初始化。

3 DDIM快速采样

预测网络给出噪声，采样器决定怎样更新状态。

训练噪声日程与采样时间选择

预测器已能回答“当前噪声是什么”，采样器还要决定“下一次去哪个时间”。以下$T = 1000$仅为教学日程。



密集链（999、998、997所在支路）：原逐步DDPM基线依次调用1000次，省略号保留其间各步。稀疏链（750、500、250所在支路）：DDIM示意子序列只调用4次，每次使用两端累计系数重新计算更新。

稀疏链含5个时间节点、4条更新边，因此调用预测器4次；CFG另计两分支预测。训练日程$\bar{\alpha}_t$保持一致。更换采样器通常可复用噪声预测网络；CFG每步两分支预测另计，合批也保留两份预测计算。

用干净预测、状态方向和新噪声更新

去噪扩散隐式模型（DDIM）采用与原训练相容的单时刻加噪边缘，重新选择采样时间和随机性。继承$a_t = \sqrt{\overline{\alpha}_t}$、$b_t = \sqrt{1-\overline{\alpha}_t}$，从当前时间t走到较早时间$r < t$。

给定当前噪声预测$\hat{\epsilon}$（可采用CFG），先得到 $\hat{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t - b_t\hat{\epsilon}}{a_t}$. 然后更新 $\boldsymbol{x}_r = a_r\hat{\boldsymbol{x}}_0 + \sqrt{b_r^2 - \rho_{t\to r}^2}\hat{\epsilon} + \rho_{t\to r}\boldsymbol{\xi}, \boldsymbol{\xi}\sim\mathcal{N}(0, \boldsymbol{I})$.

$\rho_{t\to r}$是本步新增噪声的标准差，须满足$0 \le \rho_{t\to r}^2 \le b_r^2$。常用选择为 $\rho_{t\to r} = \eta\sqrt{\frac{1-\overline{\alpha}_r}{1-\overline{\alpha}_t}}\sqrt{1-\frac{\overline{\alpha}_t}{\overline{\alpha}_r}}, 0 \le \eta \le 1.$

$\eta = 0$时各步不再抽新噪声，给定起点和条件后更新确定；初始状态仍是随机抽样。原DDIM论文中的累计α_t对应这里的$\overline{\alpha}_t$。

时间子序列上的完整生成步骤

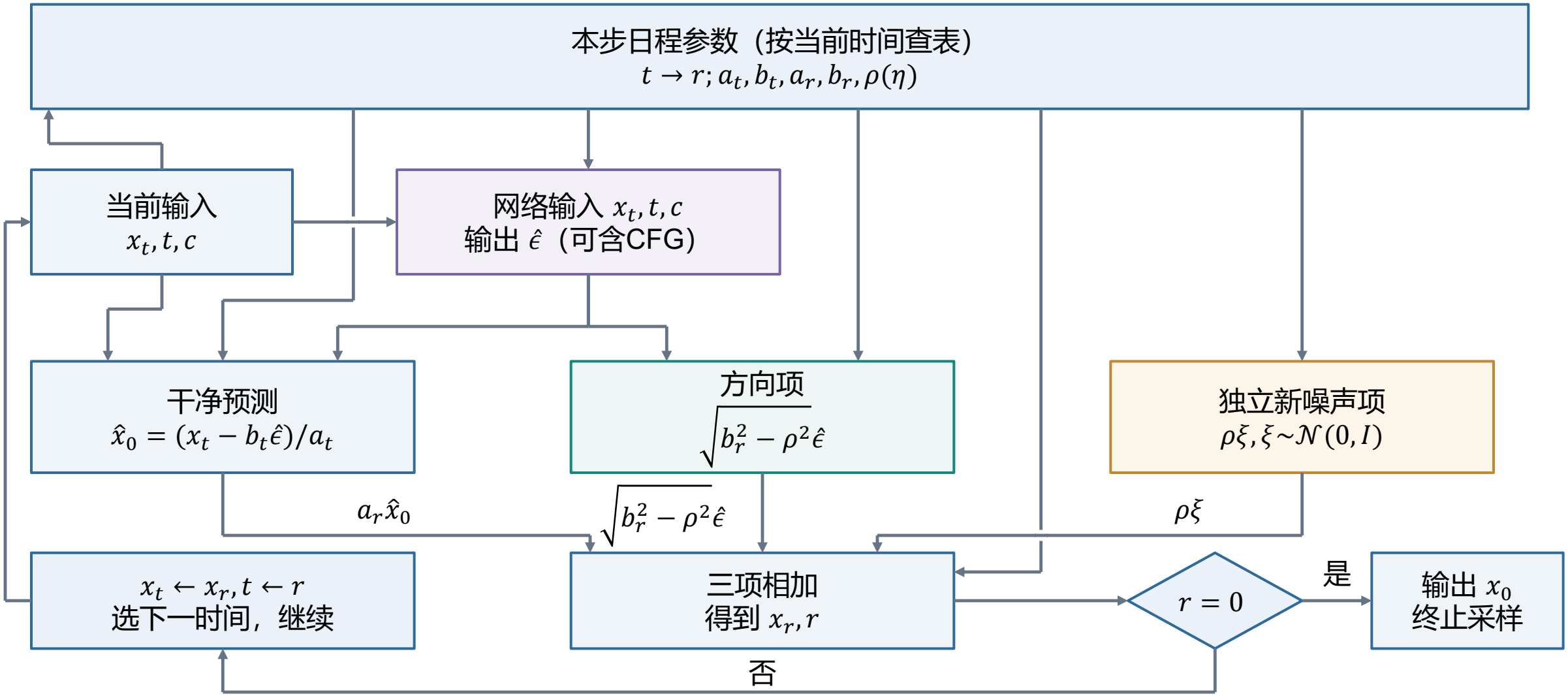
据DDIM §4.1式(12)、§4.2及附录C.1整理以下步骤。

选定$0 = \tau_0 < \tau_1 < \cdots < \tau_K = T$，生成按其逆序执行：

1. 给定条件与η，抽取$\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$，固定网络参数。
2. 对$i = K, \ldots, 1$，令$t = \tau_i, r = \tau_{i-1}$，在t处预测噪声，需要时作CFG组合。
3. 使用前页公式计算$\rho_{t\to r}$、$\hat{\boldsymbol{x}}_0$及$\boldsymbol{x}_r$，将新状态交给下一次预测。
4. 最终r=0时$a_0 = 1, b_0 = 0, \rho_{t\to 0} = 0$，输出$\hat{\boldsymbol{x}}_0$。潜在模型继续恢复尺度并解码。

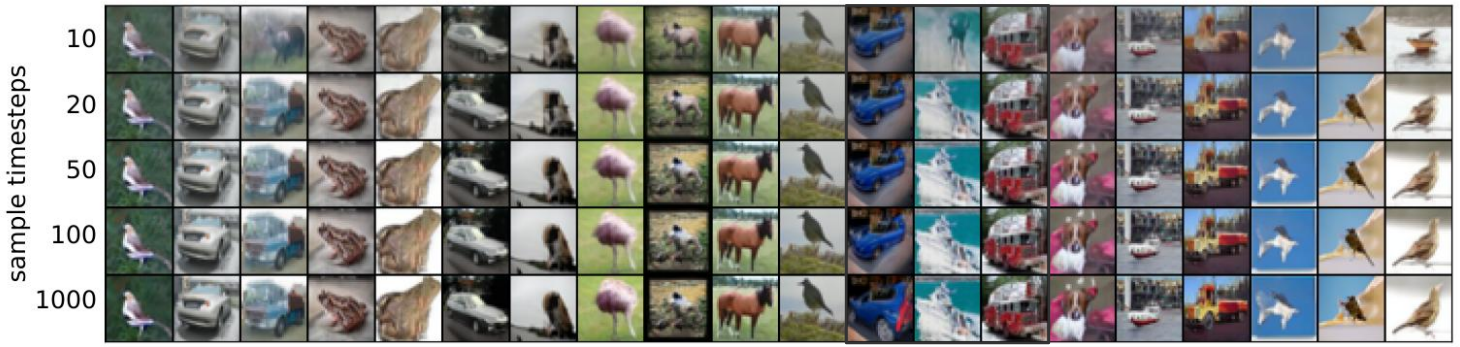
该数据端更新是确定的输出约定。论文为似然与变分论证另规定正方差末端解码分布；零方差输出不能直接代入要求非退化密度的定理。减少采样调用可降低计算，但实际质量还取决于网络、时间选择、引导和数值误差。

DDIM更新与时间循环图



同一起点、不同采样步数的生成终点

原图上部完整面板（20列 × 5行）

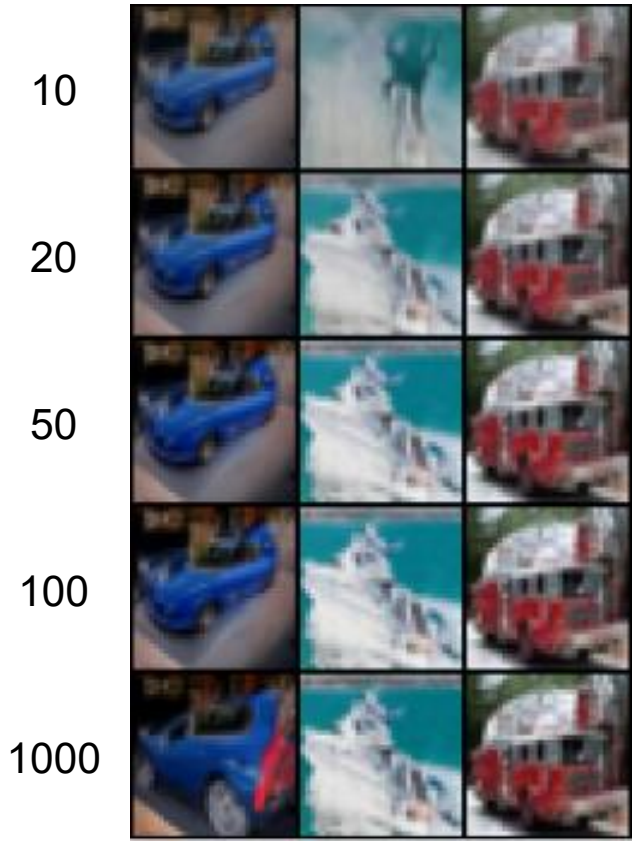


方框：本页放大的连续第12—14列

固定一列：同一初始噪声，比较不同采样总步数。

横向换列：改变初始噪声。每格均为完整生成终点。

原图局部：12—14列



来源：Song等，DDIM v4，Figure 5上部，PDF p.8；右图仅作原生裁切放大，行列不变。

完整生成系统与本讲总结

系统层次	本讲结论	可替换对象
条件与CFG	共享网络给出两种预测，用s组合	条件、引导强度
潜在表示	先学表示，再学潜在分布；尺度一致	表示空间与缩放
预测骨干	U-Net多尺度，DiT处理patch序列	预测器内部架构
DDIM采样	干净项、方向项、新噪声项联合更新	时间子序列与η

生成闭环：初始化潜在噪声；按日程把状态、时间与条件送入预测器，按需CFG；采样得到较早状态，未到0则回写状态与时间，到0后除k并解码。推理固定参数。

随堂核对：s=1使用哪种预测？η=0还剩哪种随机性？替换DiT是否必然更换DDIM？ 主线讲授在此结束。深化A保留得分与CFG证明，深化B保留DDIM推导；末页为插值拓展。

4 深化A：得分与条件引导

主线已闭合，独立证明核得分、条件平均与CFG依据。

高斯扰动核的对数梯度

给定$\boldsymbol{x}_0$，扰动核为$q_t(\boldsymbol{x} \mid \boldsymbol{x}_0) = \mathcal{N}(a_t\boldsymbol{x}_0, b_t^2\boldsymbol{I})$。在$t > 0$、$b_t > 0$处，对当前状态$\boldsymbol{x}$求梯度：

$$\log q_t(\boldsymbol{x} \mid \boldsymbol{x}_0) = C_t - \frac{\|\boldsymbol{x}-a_t\boldsymbol{x}_0\|^2}{2b_t^2},\ \nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x} \mid \boldsymbol{x}_0) = -\frac{\boldsymbol{x}-a_t\boldsymbol{x}_0}{b_t^2}.$$

C_t与当前状态无关，平方范数求导产生的2与分母中的2抵消。若当前状态由$\boldsymbol{x}_t = a_t\boldsymbol{x}_0 + b_t\boldsymbol{\epsilon}$构造，则该核梯度为$-\boldsymbol{\epsilon}/b_t$。

得分是对数密度对状态的梯度。这里的核得分给定了真实干净数据；生成时通常不知道这个条件。

边缘得分是核得分的条件平均

① 前提与边缘混合

固定观察状态x，$b_t > 0$；密度正且可微，满足求导与积分交换及梯度可积条件。

$$p_t(\boldsymbol{x}) = \int q_t(\boldsymbol{x} \mid \boldsymbol{u}) p_{\text{data}}(\boldsymbol{u}) d\boldsymbol{u}.$$

② Bayes后验权重

用$\nabla q = q \nabla \log q$，再把积分中的比值识别为来源后验：

$$p(\boldsymbol{u} \mid \boldsymbol{x}_t = \boldsymbol{x}) = \frac{q_t(\boldsymbol{x}|\boldsymbol{u}) p_{\text{data}}(\boldsymbol{u})}{p_t(\boldsymbol{x})}.$$

③ 对核得分取条件平均

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) = \int p(\boldsymbol{u} \mid \boldsymbol{x}_t = \boldsymbol{x}) \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x} \mid \boldsymbol{u}) d\boldsymbol{u} = -\frac{\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{x}_t=\boldsymbol{x}]}{b_t}.$$

同一个带噪状态可来自多个干净来源；一次噪声标签只是一个来源的实现。

平方损失学习条件平均噪声

① 有限二阶矩与正交分解

令$U=(\boldsymbol{x}_t,t,c)$，$\boldsymbol{m}(U)=\mathbb{E}[\boldsymbol{\epsilon}\mid U]$。

$$\boldsymbol{\epsilon}-\boldsymbol{f}=(\boldsymbol{\epsilon}-\boldsymbol{m})+(\boldsymbol{m}-\boldsymbol{f}),$$

$$\mathbb{E}[\langle\boldsymbol{\epsilon}-\boldsymbol{m},\boldsymbol{m}-\boldsymbol{f}\rangle\mid U]=0,$$

$$\mathbb{E}[\|\boldsymbol{\epsilon}-\boldsymbol{f}\|^2\mid U]=\mathbb{E}[\|\boldsymbol{\epsilon}-\boldsymbol{m}\|^2\mid U]+\|\boldsymbol{m}-\boldsymbol{f}\|^2.$$

第一项与f无关；第二项在$f=m$时最小。

② 同一状态，不同对象

无条件例：$x_0\sim\mathcal{N}(0,1)$，$a_t=0.8, b_t=0.6$。

在$x_t=1$：边缘得分为-1，最优噪声估计为0.6。

一次标签：$x_0=0.5,\epsilon=1$也得到$x_t=1$，但核得分为$-1/0.6$。

③ 最优估计与条件得分

$$\boldsymbol{r}_{t,c}(\boldsymbol{x})=\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}\mid c)=-\frac{\mathbb{E}[\boldsymbol{\epsilon}\mid\boldsymbol{x}_t=\boldsymbol{x},c]}{b_t}.$$

单次标签、理想回归与网络估计身份不同；学得网络只近似条件平均。

条件概率给出输入引导方向

在$p(c) > 0$、相关密度与条件概率为正的状态上，Bayes关系给出 $\log p_t(\boldsymbol{x} \mid c) = \log p_t(\boldsymbol{x}) + \log p_t(c \mid \boldsymbol{x}) - \log p(c)$. 条件$c$固定，最后一项不随状态变化。对$\boldsymbol{x}$求梯度： $\boldsymbol{r}_{t,c}(\boldsymbol{x}) = \boldsymbol{r}_t(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log p_t(c \mid \boldsymbol{x})$.

带噪分类器$h_\psi(c \mid \boldsymbol{x}, t)$可估计右侧条件概率。固定分类器参数$\psi$，对当前输入的对数概率求梯度，仍能获得引导方向；分类器须适应对应噪声等级。

以无条件得分为基准，使用权重s得到$\boldsymbol{r}_t + s\nabla \log h_\psi$；若已使用条件得分，再添加权重$w$，在理想一致模型下对应$s = 1 + w$。不同基准需要不同系数。

对数概率梯度需要归一化项

若分类器输出logit$\ell_j(\boldsymbol{x})$，归一化概率为$h_j = \exp \ell_j / \sum_k \exp \ell_k$。对指定类别c： $\nabla \log h_c = \nabla \ell_c - \sum_j h_j \nabla \ell_j$. 减去的归一化项通常依赖输入，所以单个logit的梯度通常不能替代对数概率梯度。

设$h(c = 1 \mid x, t) = sigmoid(2x)$。在$x = 0$处概率为0.5，$\partial_x \log h = 2(1 - 0.5) = 1$，logit梯度却是2。

固定时刻的加权场与CFG边界

① 固定时间：理想场的解释

固定同一t，密度精确、相容且可归一化，$b_t > 0$。

$\boldsymbol{r}_t + s\nabla \log p_t(c \mid \boldsymbol{x}) = \nabla \log[p_t(\boldsymbol{x}) p_t(c \mid \boldsymbol{x})^s]$.

以理想无条件/条件回归$\boldsymbol{\epsilon}_{\mathrm{u}}^*, \boldsymbol{\epsilon}_{\mathrm{c}}^*$代入得分关系：

$$\boldsymbol{r}_t + s(\boldsymbol{r}_{t,c} - \boldsymbol{r}_t) = -\frac{\boldsymbol{\epsilon}_{\mathrm{u}}^* + s(\boldsymbol{\epsilon}_{\mathrm{c}}^* - \boldsymbol{\epsilon}_{\mathrm{u}}^*)}{b_t}.$$

$s = 1$恢复原条件得分；实际CFG用网络预测替代理想回归。

② 整条采样：仍需跨时间条件

各时刻的加权密度未必是同一前向过程的一致边缘族。网络近似与离散更新还会引入误差。 因此，固定时刻的得分恒等式不提供整条实际采样精确生成加权分布或原条件分布的保证。

5 深化B：DDIM构造与更新

独立证明边缘保持、模型替代、KL与合法时间跳跃。

单时刻边缘没有确定整条路径

训练噪声回归在每次抽取的t上使用$(\boldsymbol{x}_0, \boldsymbol{x}_t, \boldsymbol{\epsilon})$。只要 $q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(a_t\boldsymbol{x}_0, b_t^2\boldsymbol{I})$ 保持不变，该时间的噪声回归数据分布就不变；不同时间状态之间的联合关系仍可改变。

固定$\boldsymbol{x}_0$，考虑两个时刻的标准化残差$\boldsymbol{e}_t, \boldsymbol{e}_r$。若二者独立，或令$\boldsymbol{e}_r = \boldsymbol{e}_t$，每个残差单独都为标准高斯，但跨时刻相关性不同。对应$\boldsymbol{x}_j = a_j\boldsymbol{x}_0 + b_j\boldsymbol{e}_j$具有相同单时刻边缘、不同路径联合。

DDIM由终端边缘和较晚到较早的条件分布构造一种路径族： $q_\rho(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = q_T(\boldsymbol{x}_T \mid \boldsymbol{x}_0)\prod_{t=2}^{T} q_\rho\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)$. 这是一种按反向条件写出的推断路径；用Bayes改写为前向时，通常仍依赖$\boldsymbol{x}_0$，所以可为非马尔可夫过程。

构造保持目标边缘的高斯连接

① 前提与高斯连接

$0 < r < t$，$b_t > 0$，固定真实x_0；旧残差$e = (x_t - a_t x_0)/b_t$与新ξ独立，均为标准高斯。

$$\boldsymbol{x}_r = a_r \boldsymbol{x}_0 + \sqrt{b_r^2 - \rho^2}\,\boldsymbol{e} + \rho\boldsymbol{\xi}, 0 \le \rho^2 \le b_r^2.$$

$\breve{d}_{t,r}$

旧残差的标准差为$d_{t,r}$，新独立噪声的标准差为ρ。

② 均值与协方差验证

$$\mathbb{E}[\boldsymbol{x}_r \mid \boldsymbol{x}_0] = a_r \boldsymbol{x}_0,$$

$$Cov(\boldsymbol{x}_r \mid \boldsymbol{x}_0) = (b_r^2 - \rho^2)\boldsymbol{I} + \rho^2 \boldsymbol{I} = b_r^2 \boldsymbol{I}.$$

独立性使交叉协方差为0；方差之和恢复b_r^2。

③ 结论与退化边界

独立高斯的线性组合仍是高斯，故保持目标边缘；从终端逆向归纳全部所选时刻。 给定x_t, x_0，条件均值为构造式前两项，协方差$\rho^2 I$。$\rho = 0$定义确定映射，条件分布退化。

用模型预测替代未知干净数据

生成时没有真实$\boldsymbol{x}_0$。从加噪恒等式可解出 $\boldsymbol{x}_0 = (\boldsymbol{x}_t - b_t\boldsymbol{\epsilon})/a_t$. 选用网络预测$\hat{\boldsymbol{\epsilon}}$定义$\hat{\boldsymbol{x}}_0 = (\boldsymbol{x}_t - b_t\hat{\boldsymbol{\epsilon}})/a_t$。

将它替入前页条件均值，并保留模型选定的随机项：

$$\begin{aligned}\boldsymbol{x}_r &= a_r\hat{\boldsymbol{x}}_0 + \sqrt{b_r^2 - \rho^2}\,\frac{\boldsymbol{x}_t - a_t\hat{\boldsymbol{x}}_0}{b_t} + \rho\boldsymbol{\xi} \\ &= a_r\hat{\boldsymbol{x}}_0 + \sqrt{b_r^2 - \rho^2}\,\hat{\boldsymbol{\epsilon}} + \rho\boldsymbol{\xi}.\end{aligned}$$

第二行使用干净预测定义，因而分式等于预测噪声。真实路径构造中的等式与生成模型中的替代分别成立；替代误差及引导会影响实际状态分布。

若$\rho = 0$，初始状态和条件固定后每步都是确定计算。这一结论说明新随机量被移除，不保证恢复某个真实数据或准确逆转任意已实现路径。

正方差目标对应带权噪声误差

① 正方差条件与均值差

中间时刻$\rho > 0$，两高斯使用同一$\rho^2 I$，$a_t, b_t > 0$；记$d_{t,r} = \sqrt{b_r^2 - \rho^2}$。

$$\boldsymbol{m}(\boldsymbol{u}) = \frac{d_{t,r}}{b_t}\boldsymbol{x}_t + \left(a_r - \frac{d_{t,r}a_t}{b_t}\right)\boldsymbol{u}.$$

训练代入$u = x_0$，模型代入$u = \hat{x}_0$。共同状态项相消：

$$\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0 = \frac{b_t}{a_t}(\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}),$$

$$\Delta m = \left(\frac{a_r b_t}{a_t} - d_{t,r}\right)(\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}).$$

② 同协方差KL

均值差平方除以$2\rho^2$：

$$D_{\mathrm{KL}}(q_\rho \| p_\theta) = \frac{\left(a_r b_t / a_t - d_{t,r}\right)^2}{2\rho^2} \cdot \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\|^2.$$

固定日程与方差给出正的时间权重。

③ 目标与适用边界

负ELBO还含终端项与正方差末步似然；相同边缘支持原数据估计。各时刻独立理想模型可分别最小化，共享网络须权衡。 无权目标是替代目标；$\rho = 0$不满足本KL条件，确定性采样另行定义。

跨越多个时间步仍须重建合法连接

① 稀疏连接的合法范围

每对$t > r > 0$保留原两端累计系数，调用次数由子序列决定。设$0 < \bar{\alpha}_t < \bar{\alpha}_r < 1$，$0 \le \eta \le 1$：

$$\rho^2 = \eta^2 \frac{1-\bar{\alpha}_r}{1-\bar{\alpha}_t}\left(1-\frac{\bar{\alpha}_t}{\bar{\alpha}_r}\right).$$

由$\bar{\alpha}_t/\bar{\alpha}_r \ge \bar{\alpha}_t$，得到

$$0 \le \frac{1-\bar{\alpha}_t/\bar{\alpha}_r}{1-\bar{\alpha}_t} \le 1,$$

故$0 \le \rho^2 \le 1-\bar{\alpha}_r = b_r^2$。

② 相邻DDPM特例

令$r = t-1$、$\eta = 1$，用$\bar{\alpha}_t = \alpha_t\bar{\alpha}_{t-1}$：

$$\rho^2 = \frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} = \tilde{\beta}_t$$

这是DDPM的后验方差选择；GM05例用β_t，两种设置须区分。

③ 数据端单独处理

$r = 0$时$b_0 = \rho = 0$，算法输出干净预测。论文变分论证另保留正方差末步解码以支持密度；本课的数据端显示更新不冒充该模型的完整随机抽样。

同一状态上的确定性与随机更新

采用一维教学构造：$\overline{\alpha}_t = 0.64, \overline{\alpha}_r = 0.81$，当前$x_t = 1.1$，预测$\hat{\epsilon} = 0.2$。于是 $\hat{x}_0 = (1.1 - 0.6\times0.2)/0.8 = 1.225.$

当$\eta = 0$，$\rho = 0$，得到 $x_r = 0.9\times1.225 + \sqrt{0.19}\times0.2 \approx 1.189678.$

当$\eta = 1$，$\rho^2 = 0.19(1 - 0.64/0.81)/0.36 \approx 0.110768$，方向系数$\sqrt{0.19 - \rho^2} \approx 0.281481$。若新噪声$\xi = -0.5$，则 $x_r = 1.1025 + 0.281481\times0.2 + 0.332819\times(-0.5) \approx 0.992387.$

两种更新使用同一状态和同一预测，只改变随机性选择。它们不是将同一个预测结果简单加噪：方向项系数也随ρ改变。这是人工构造的运算核对，不是质量实验。

附录：确定性DDIM的初始噪声插值

图像按初始噪声的插值位置从左到右排列。中间起点取两端噪声的球面插值。

每个起点运行50步确定性DDIM，输出最终像素图像x_0。

连续改变的是x_T的插值位置，不是去噪时间t。这些面板不是x_T、x_r、x_0三种状态，也不是潜在特征的可视化；不据此证明任意语义都线性变化。

来源：Song等，DDIM，Fig.6顶行（arXiv:2010.02502v4，p.8）；方法见附录D.5。原论文结果，非本课实验。

