08 生成模型评价与前沿问题

根据任务与证据理解生成系统

学习目标

本讲问题：怎样判断一个生成模型真正有效?

- 按任务区分质量、覆盖、条件、成本与记忆，选择相应证据。
- 解释FID和IS观察什么、如何计算，以及分数不能保证什么。
- 阅读模型比较，核对数据、观察器和生成协议，给出有限结论。

知识路线：评价问题 → 指标与盲区 → 多种证据 → 比较协议与结论边界。

扩展模块：少步生成、表示空间、离散状态与视频结构；用同一评价框架审查技术变化。

1 任务与指标使用

区分评价问题，建立FID与IS的使用接口。

不同评价问题需要不同观察

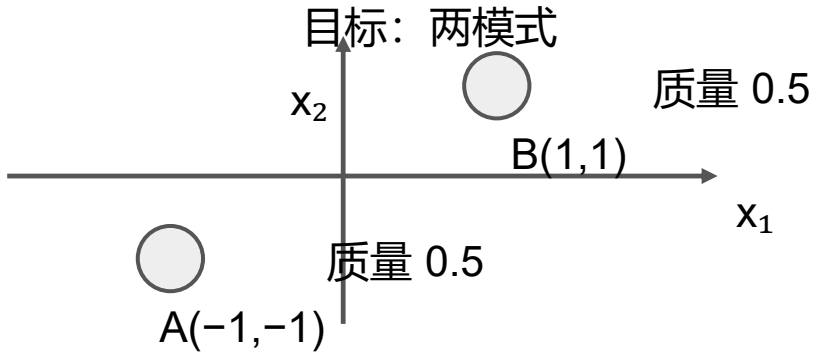
给定“红球在蓝盒左侧”的生成任务：清晰、对象齐全、关系正确和生成快，是同一个问题吗？先确定要判断的性质，再选择数据与观察。

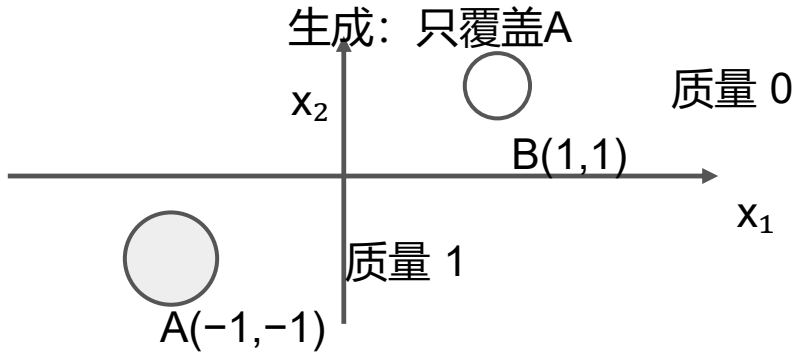
评价问题	主要观察
模型怎样分配观测概率?	同一数据定义下的独立数据似然或有条件的估计
输出是否具有所需局部质量?	按明确方式抽取的样本及任务相关判断
是否表达目标的不同情况?	模式、类别及其他任务属性的覆盖
是否遵循给定条件?	条件与对应输出的匹配
需要多少生成资源?	明确运行条件和测量范围的成本

这是评价问题的对照，后续流程仍需具体定义。对连续数据使用密度，对离散数据使用概率质量；变分下界和精确似然具有不同身份。直接比较数值前，必须说明观测单位和预处理。

局部正确可以伴随覆盖不足

人工构造目标：二维点$A = (-1,-1)$、$B = (1,1)$各占一半。模型只输出A。每个生成点都位于真实模式，B却完全缺失。





合理位置与目标覆盖需要分别检查。接下来先看FID如何将两组无法逐张对应的图像转成可比较的统计。

独立数据上的较高似然支持指定概率目标下的表现，视觉质量和任务符合度仍须另有证据。高维情形中，似然与样本视觉判断可能给出不同结论，评价应回到使用目的。

FID比较指定特征的一、二阶统计

两组图像不能逐张对齐，如何比较？**FID（Fréchet Inception Distance）**先用同一固定特征提取器，再比较均值与协方差。



r：参考　g：生成；均值=中心，协方差=波动与相关

$$FID = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|^2 + tr[\boldsymbol{C}_r + \boldsymbol{C}_g - 2(\boldsymbol{C}_r^{1/2}\boldsymbol{C}_g\boldsymbol{C}_r^{1/2})^{1/2}].$$

第一行为中心差；第二行为波动尺度与相关结构差。迹是对角和，平方根是半正定矩阵平方根。

同一协议下较低值表示特征矩更接近，不能保证条件正确、覆盖充分或无复制。样本统计有误差，特征需有有限二阶矩。

IS比较单图预测与生成集合的类别边缘

换一个观察器：如果只读取生成图像的分类预测，会保留哪些信息?

固定分类器给出$q(y \mid \boldsymbol{x})$，生成样本$\boldsymbol{x} \sim P_g$。对生成集合平均得到类别边缘 $q(y) = E_{\boldsymbol{x} \sim P_g} q(y \mid \boldsymbol{x})$.

Inception Score（IS）定义为 $IS = \exp\left(E_{P_g} D_{\mathrm{KL}}(q(y \mid \boldsymbol{x}) \| q(y))\right)$. 有限类别下，KL为 $\sum_y q(y \mid \boldsymbol{x}) \ln[q(y \mid \boldsymbol{x}) / q(y)]$，零项取连续约定，自然对数对应指数还原。

IS倾向于单图预测集中、整组预测类别分散。两个极端例分别为：

- 所有样本都被确定预测为同一类别时，IS=1
- 在两个类别间均衡且各样本预测确定时，IS=2

IS不直接比较真实图像分布，也不能辨认同一类别内是否反复复制一张图。

分类器置信度描述预测性质。真实性和条件符合度需其他证据。

FID / IS比较必须同时固定样本数、数据与处理实现。

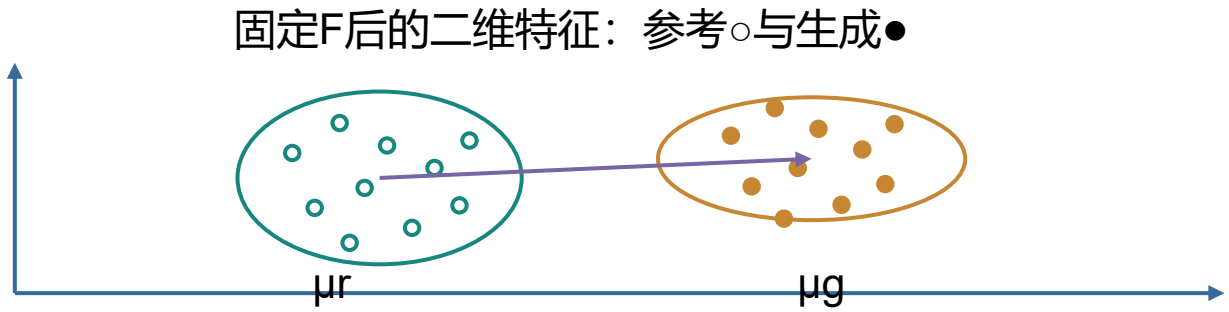
分类概率也有盲区；深度部分将逐项拆开计算。

2 指标机制与统计盲区

特征矩与分类预测的机制；可跳至第3节。

图像经过固定观察器后形成矩统计

设参考图像与生成图像分别有$n_r, n_g \geq 2$个。相同评价定义包含各自已说明的预处理，再用固定特征映射F得到$\boldsymbol{f}_i \in \mathbb{R}^d$。



中心差 → 均值项

椭圆示意波动与相关 → 协方差项

图为特征统计示意；椭圆不是精确的真实支持边界

对任一组n个特征，样本均值和无偏样本协方差为

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_i \boldsymbol{f}_i,$$

$$\hat{\boldsymbol{C}} = \frac{1}{n-1}\sum_i (\boldsymbol{f}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{f}_i - \hat{\boldsymbol{\mu}})^{\top}.$$

协方差刻画分量的共同波动。有限样本的协方差估计无偏，不推出非线性FID无偏。

均值项与协方差项比较不同变化

样本矩已经得到，协方差差异如何进入FID?

均值差平方比较特征中心。协方差项在用高斯近似特征分布时，衡量波动尺度与相关结构的差别。这里引用高斯间平方Wasserstein距离的结果，得到FID公式。

半正定矩阵$\boldsymbol{A}$可写为

$$\begin{aligned}\boldsymbol{A} &= \boldsymbol{U}diag(\lambda_j)\boldsymbol{U}^{\top}, \lambda_j \geq 0,\\ \boldsymbol{A}^{1/2} &= \boldsymbol{U}diag(\sqrt{\lambda_j})\boldsymbol{U}^{\top}.\end{aligned}$$

这不是逐元素开平方。采用 $\boldsymbol{A} = \boldsymbol{C}_r^{1/2}\boldsymbol{C}_g\boldsymbol{C}_r^{1/2}$ 时，它仍对称半正定，可以使用该定义。

原论文常写$tr[(\boldsymbol{C}_r\boldsymbol{C}_g)^{1/2}]$。当$\boldsymbol{C}_r$正定时，乘积与上述A相似，因此对应主平方根的迹相同。本讲对可能退化的协方差直接采用对称表达，避免把非对称乘积当作对称矩阵逐项操作。

下一页把矩阵退到一维，直接复算两项贡献。

一维特征例核对均值和标准差

标量化简：一维协方差就是方差，设标准差$\sigma_r, \sigma_g \geq 0$。

$$\sigma_r^2 + \sigma_g^2 - 2\sqrt{\sigma_r^2 \sigma_g^2} = (\sigma_r - \sigma_g)^2.$$

$$FID_{1D} = (\mu_r - \mu_g)^2 + (\sigma_r - \sigma_g)^2.$$

参数代入：参考均值0、方差1；生成均值1、方差4，即标准差2。

结果变化：

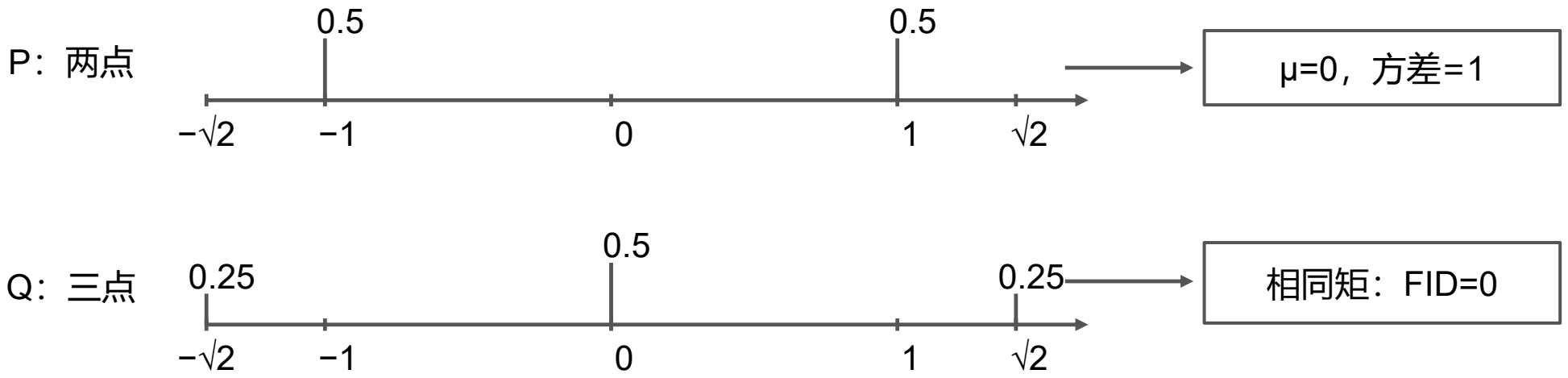
- 原设定：均值项1，标准差项1，FID=2。
- 仅将生成方差改为1：FID=1。
- 再将生成均值改为0：FID=0。

这是固定特征的人工计算，不是实际图像模型结果。矩匹配得到0，是否意味着分布相同？

相同前两阶矩仍可隐藏分布差异

在同一个一维特征空间构造： $P(f=-1)=P(f=1)=\frac{1}{2}, Q(f=-\sqrt{2})=Q(f=\sqrt{2})=\frac{1}{4}, Q(f=0)=\frac{1}{2}.$

两者均值为0。P的二阶矩为1；Q的二阶矩为 $\frac{1}{4}(2)+\frac{1}{2}(0)+\frac{1}{4}(2)=1$。 均值为0时，二阶矩等于方差，所以两者方差均为1，FID统计值为0。

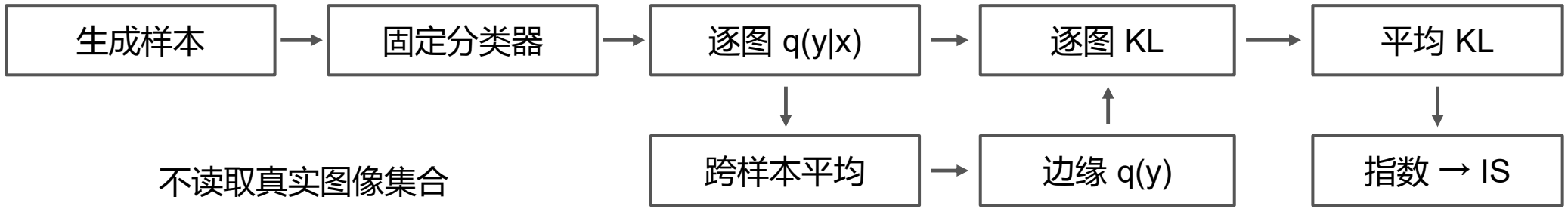


P与Q在0处的质量分别为0与1/2，分布明显不同。固定观察器还可能丢失其他图像信息；因此低FID只支持指定特征矩接近。条件正确、细粒度覆盖和无记忆分别需要额外证据。

单图预测与集合平均具有不同身份

固定分类器预测$q(y \mid \boldsymbol{x})$，并对生成分布平均： $q(y) = E_{\boldsymbol{x}\sim P_g} q(y \mid \boldsymbol{x})$. 取K个有限类别，

$\sum_y q(y \mid \boldsymbol{x}) = 1$。IS使用的核心量为 $A = E_{P_g} \sum_y q(y \mid \boldsymbol{x}) \ln \frac{q(y|\boldsymbol{x})}{q(y)}$.



若$q(y) = 0$，非负性意味着该类别预测在生成分布下几乎处处为0，按零项约定处理。这里的Y是分类器输出类别，不直接表示人类判断的真实性或文本匹配。

把平均KL展开，就能分清单图确定性与集合类别变化。

平均KL展开为边缘熵减平均条件熵

把对数比拆开，对有限类别交换求和与定义良好的期望：

$$
\begin{aligned}
A \quad &= E_{\boldsymbol{x}} \sum_y q(y \mid \boldsymbol{x}) \ln q(y \mid \boldsymbol{x}) \\
&\quad - \sum_y E_{\boldsymbol{x}}\, q(y \mid \boldsymbol{x}) \ln q(y) \\
&= -E_{\boldsymbol{x}} H(q(\cdot \mid \boldsymbol{x})) + H(q).
\end{aligned}
$$

熵$H(q) = -\sum_y q(y) \ln q(y)$刻画类别分布的不确定性。

第一项来自单图平均熵，第二项由类别边缘得到：

$$IS = \exp(A), A = H(q) - E_{\boldsymbol{x}} H(q(\cdot \mid \boldsymbol{x})).$$

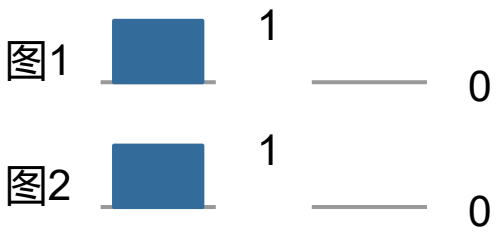
类别边缘更分散可提高第一项，单图预测更集中可降低被减去的项；这两种类别统计仍不能辨认类内变化。

对K类有$1 \le IS \le K$；上界需类别边缘均匀且每张图预测确定。

单类确定与处处均匀都可得到IS为1

采用两类预测，分别考察三种人工构造。

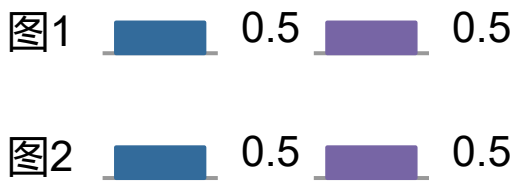
全部确定为类1



q(y)=(1.0, 0.0)

IS = 1

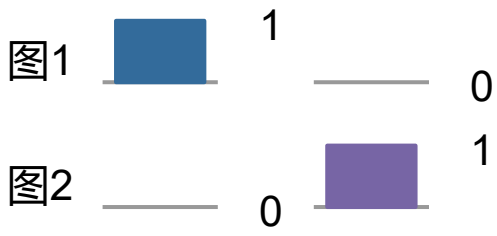
每张图都不确定



q(y)=(0.5, 0.5)

IS = 1

单图确定、类别均衡



q(y)=(0.5, 0.5)

IS = 2

两类顺序：深蓝=类1，紫=类2；每组显示两张图的预测概率

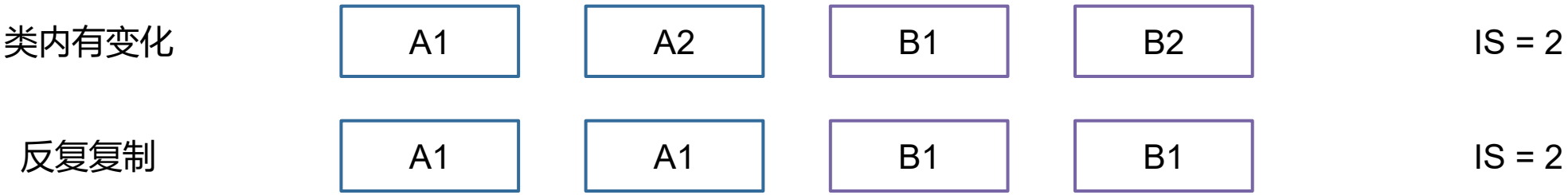
前两种都得到1，但原因不同。

第一种没有预测类别变化，第二种每张图都不确定。

相同分数不意味着相同预测行为。

类别统计不能辨认类内复制

人工构造：A类→类1、B类→类2，分类预测确定，两类各半。



颜色=类别；数字=类内样本身份；课程构造

两行分类统计相同，IS均为2；下行只复制A1/B1，类内变化更少。

软概率对照：一半预测$(0.9, 0.1)$，另一半反转；类别边缘仍均匀。

$$H(0.9,0.1) = -0.9\ln 0.9 - 0.1\ln 0.1 \approx 0.325083,$$

$$IS = \exp(\ln 2 - 0.325083) \approx 1.444935.$$

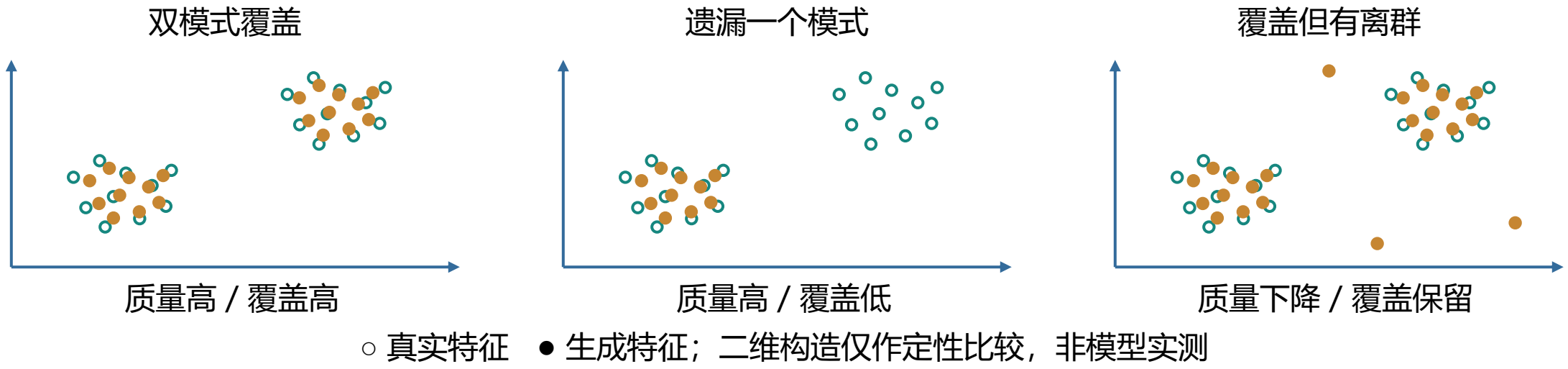
样本重复不等于复制训练实例；训练复制须核对训练来源，不能只凭重复判断。

3 评价证据、比较协议与结论边界

先选择评价证据，再固定协议，判断结论边界。

生成Precision与Recall分开估计质量和覆盖

FID / IS的盲区说明：需要把“生成点是否合理”与“真实变化是否覆盖”分开观察。



生成Precision：生成特征中，落入真实样本邻域所估计支持区域的比例。

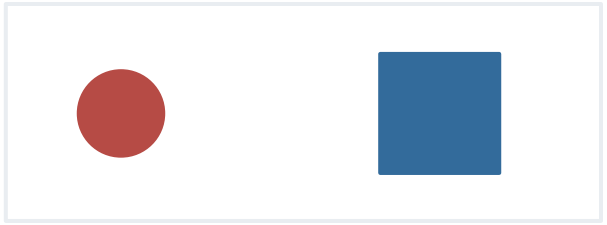
生成Recall：真实特征中，落入生成样本邻域所估计支持区域的比例。

Kynkäänniemi等（NeurIPS 2019）用特征空间的近邻球并集估计区域。结果依赖特征、样本数及邻域参数k；这里不是分类任务的查准率 / 查全率。

条件评价要检查对象、属性绑定与关系

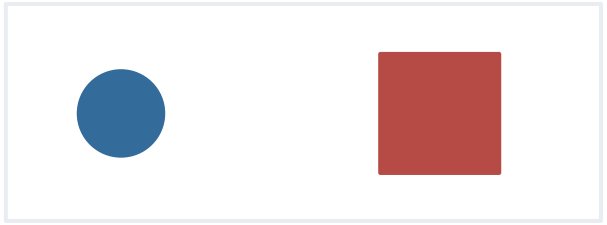
提示：“一只红色小球位于蓝色盒子的左侧。”图像清晰仍可能违反条件。

满足条件



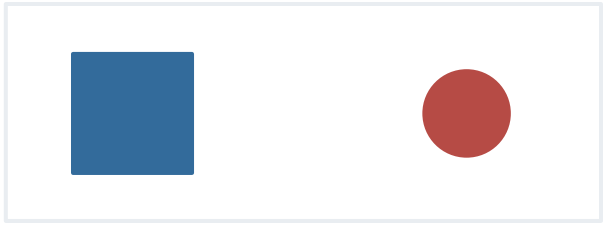
对象✓ 属性✓ 关系✓

属性绑定互换



对象✓ 属性× 关系✓

左右关系反向



对象✓ 属性✓ 关系×

依次检查：①球与盒都存在；②红色绑定球、蓝色绑定盒；③球位于盒左侧。

图文特征相似度提供整体语义线索，但高相似度不保证每项约束成立。细粒度基准可分别检查属性和关系，例如T2I-CompBench / CompBench++。

先明确任务约束，再选择检测、问答、规则或人工检查；检查器自身也会出错。

统计、条件、人工与任务评价各有职责

条件可以逐项检查，感知偏好与实际效用还需要其他观察。

证据	适合回答	主要限制
分布统计	集合级统计是否接近?	可能隐藏个别错误
条件检查	指定约束是否满足?	依赖约束与检查器
成对人工偏好	按指定标准更喜欢哪一个?	评价者差异与抽样偏差
任务测试	输出能否完成实际任务?	结论依赖任务与使用环境

成对比较示例：对同一提示匿名展示A/B，统一询问“哪张更符合对象、颜色与左右关系？”，随机左右顺序，允许平局，记录样本与评价者数量及分歧。

盲评减少模型名影响；统一问题固定标准；随机化减少位置效应。偏好不等于概率建模质量。

按评价目标选择证据并识别遗漏

读论文先定位目标，再检查对应证据；多个分数也可能观察同一种性质。

目标	代表证据	不能单独保证
概率建模	同定义独立数据似然 / 适用下界	主观真实感
分布相似	FID；KID等特征分布指标	细粒度覆盖
质量与覆盖	生成Precision / Recall	全部任务属性
条件符合	约束检查、相似度、专项基准	全部事实与关系
人类感知	标准化人工评价 / 偏好	概率建模质量
成本	NFE、延迟、吞吐、显存与训练预算	生成质量
记忆与复制	训练实例近邻、来源与专门审计	绝对无记忆
视频质量	视频分布、跨帧及专项评价	全部语义与物理性质

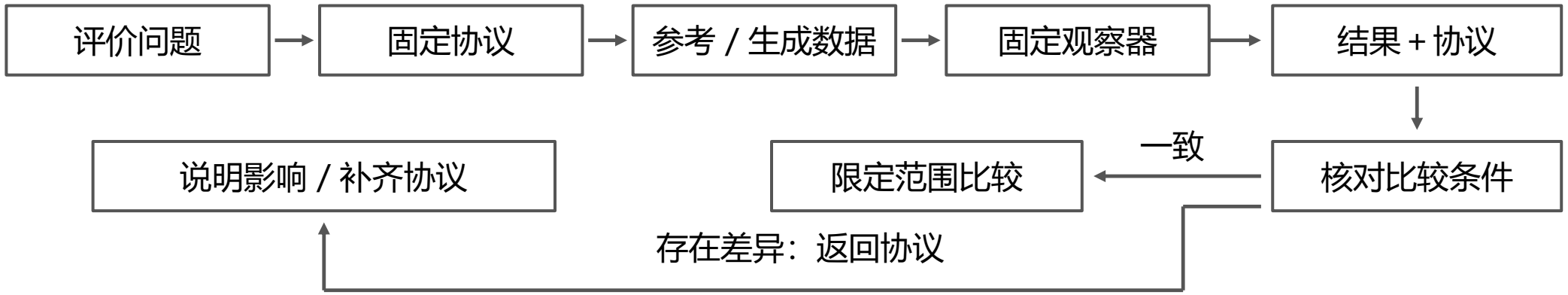
选好证据后还不能直接比较：数据、观察器和生成设置必须可恢复。

同名指标仍需相同评价定义

评价结论应附带可恢复的协议，至少说明：

- 说明参考集合、训练 / 独立划分、分辨率、样本数和样本选择方式。
- 说明特征或分类器的实现、权重、缩放、压缩、数值范围及统计估计方法。
- 说明模型版本、条件集合、引导系数、采样器、时间设置及随机性。
- 报告硬件、批量、精度、预热及测量范围，并分别报告训练与推理预算。

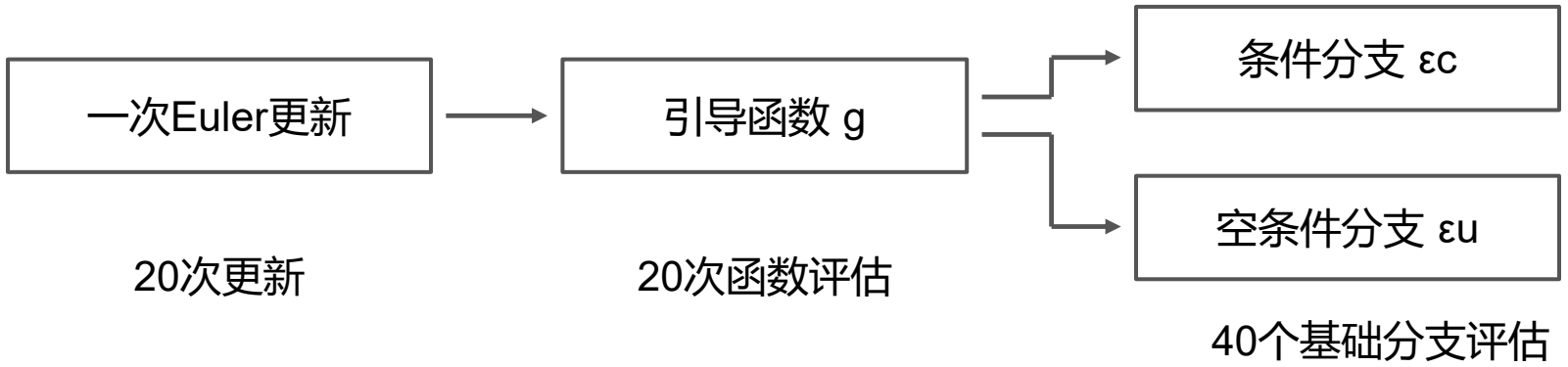
若两个报告分别对5千与5万生成样本计算FID，或使用不同缩放实现，单个数值差不能直接归因于模型。先识别协议差异，再决定补齐条件、重算或限定结论；本讲只讲判断原则，不安排评测实验。



步数与函数评估次数分别计数

步数是求解器执行的状态更新次数。**NFE**是所声明函数的评估次数；一次更新可能需要多次评估，必须说明统计的是速度场、引导后的预测函数，还是基础网络分支。

人工计数例：20次Euler更新，每步评估一次引导函数；CFG每次用条件与空条件两种预测。



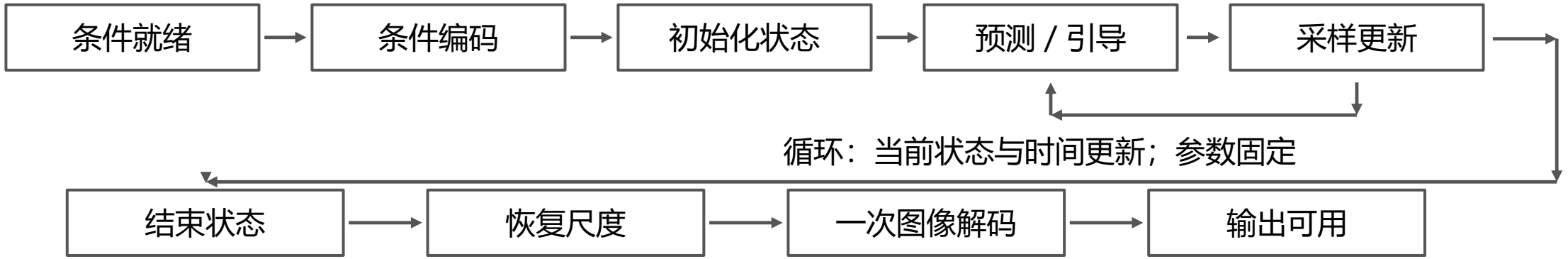
可合批：20次执行
每次含两个分支

两条件合批可以减少执行调用次数；计算量与延迟仍取决于实际实现。

采用每步两次场评估的20步求解器时，相应计数为40次引导函数评估、80个基础分支评估。这里仅解释声明清楚的计数，不比较生成质量，也不把合批次数当成计算量相同的证明。

端到端延迟包含生成链的实际范围

端到端延迟指从规定输入开始，到规定输出可用为止的耗时。潜在条件生成可能包括：

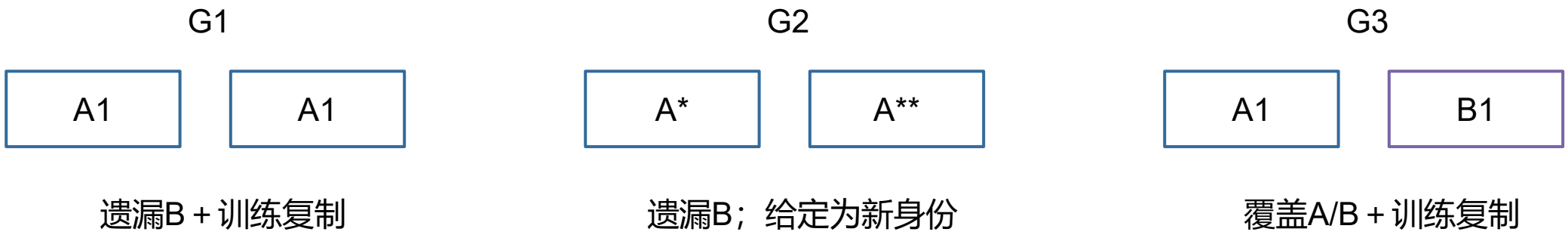


相同NFE的网络规模、分辨率、批量和精度可以不同；条件编码和解码的成本也可不同。延迟与吞吐分别回答一次请求需等待多久、单位时间能处理多少样本，二者随批量与调度改变。

报告需要说明是否计入条件编码、解码、数据传输与预热，以及硬件和并发条件。训练成本另包括数据、计算与预训练 / 蒸馏来源；低推理调用数不能单独证明总成本较低。

样本展示与独立数据提供不同证据

协议一致后，还需区分“覆盖了哪些模式”和“是否复制训练实例”。共同教学设定：训练参考为A1、A2、B1、B2，独立测试集也含A/B两类。



- 按明确规则抽样并观察类别：检查模式遗漏；精选一张A不能揭示B是否缺失。
- 与训练参考及来源核对：检查复制线索；覆盖两类不排除逐样本复制。
- 在独立数据上按同一目标评价：支持该目标下的泛化；训练—测试差距不能单独定位记忆机制。

这三种证据互补，任何单项都不等于完整证明。下一页检验近邻线索对小变换的敏感性。

近邻检查能发现线索，也有观察盲区

在可访问训练参考集合中，按指定距离检索生成样本的近邻。下面用2×2二值图说明小平移如何改变像素近邻。



二维二值图构造；平方像素距离：G到A为0、到B为2，平移后反转

检索线索 → 核对来源；未命中 → 限于当前距离与可访问参考集合

小平移、裁剪或其他变换可以改变像素距离和最近邻身份；换成感知距离可改变敏感性，但仍不是完整的无记忆证明。模型也可能只覆盖很小的总体区域，却不直接输出某个训练实例。

低FID、较高IS或近邻未命中都不单独证明无记忆；结论需注明距离、参考覆盖、预处理与展示数量。

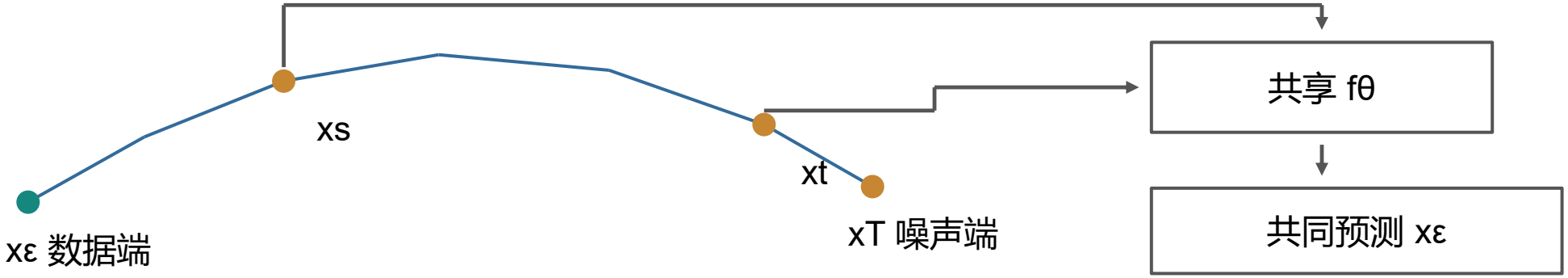
用相同框架审查前沿：改变了什么，质量与成本怎样验证？

4 少步生成改变什么

独立选讲：采样规则、直接映射与平均速度。

Consistency让同轨迹状态预测同一终点

少步改进如何验证？先问改变的是采样器，还是学得的映射。GM06的DDIM在既有预测接口上选择时间子序列；Consistency Models学习轨迹到终点的映射。



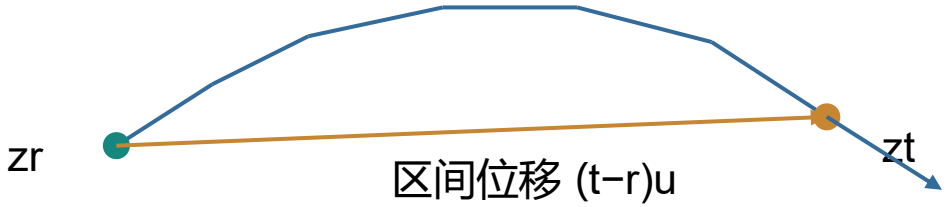
曲线：同一PF ODE轨迹；箭头：不同状态的终点预测

理想一致性函数沿同一轨迹满足 $f(\boldsymbol{x}_t, t) = f(\boldsymbol{x}_s, s) = \boldsymbol{x}_\epsilon$. 关键是不同状态的输出一致；普通的一步映射本身不包含这个关系。数据端以小正时间ϵ近似。

Song等（ICML 2023）分别讨论预训练扩散蒸馏与独立训练；支持一步，也可多步使用。比较时同时报告训练来源、质量和端到端成本。

MeanFlow学习区间位移对应的平均速度

沿同一ODE轨迹$d\boldsymbol{z}_t/dt = \boldsymbol{v}(\boldsymbol{z}_t, t)$，固定$r < t$。本文0为数据、1为噪声，生成方向与GM07相反。



定义：r → t

生成：t → r

$$\boldsymbol{u}(\boldsymbol{z}_t, r, t) = \frac{\boldsymbol{z}_t - \boldsymbol{z}_r}{t-r} = \frac{1}{t-r}\int_r^t \boldsymbol{v}(\boldsymbol{z}_\tau, \tau)d\tau.$$

固定r沿轨迹求总导数，必须包含状态随t的变化：

$$\boldsymbol{u} + (t-r)\frac{d\boldsymbol{u}}{dt} = \boldsymbol{v}.$$

训练由该恒等式构造目标；学得平均速度后，直接更新

$$\hat{\boldsymbol{z}}_r = \boldsymbol{z}_t - (t-r)\boldsymbol{u}_\theta(\boldsymbol{z}_t, r, t).$$

取t=1、r=0得到一步采样；瞬时切向不能代替整段平均。

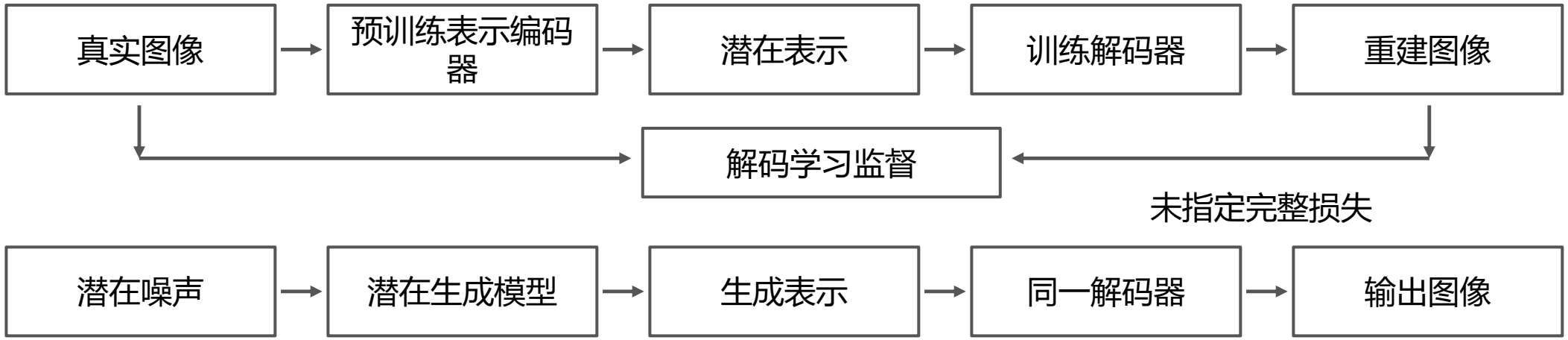
5 表示模块与RAE

独立选讲：表示编码、解码与潜在生成。

表示编码器、解码器与潜在生成分别判断

除了改变采样过程，还可以改变生成模型工作的表示空间。

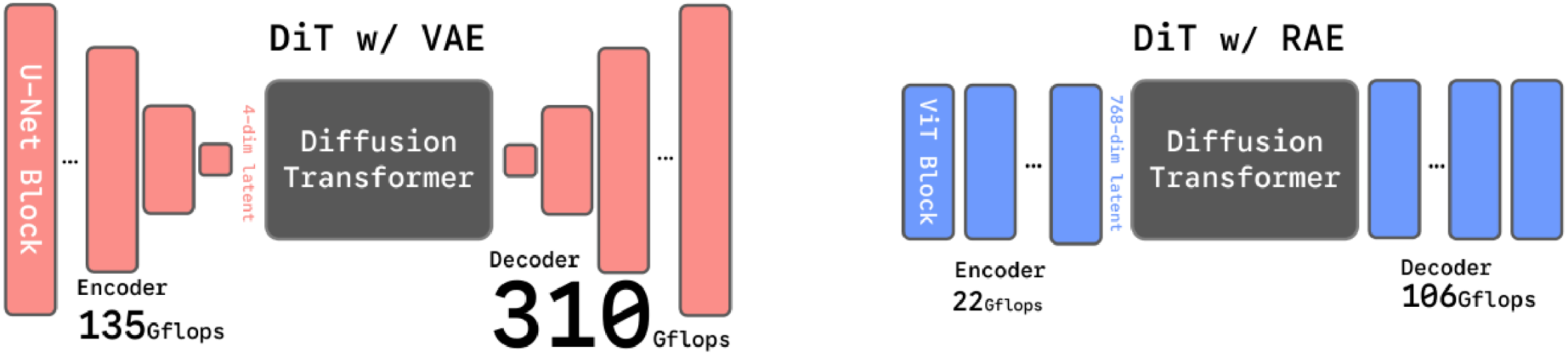
GM06已把表示学习和潜在生成分开。**Representation Autoencoder（RAE）**案例将预训练表示编码器与训练得到的解码器组合，再在相应潜在空间建立生成模型。



RAE以DINO、SigLIP、MAE等预训练表示作为候选，讨论高维潜在空间对生成模型的适配困难。接着用原结构图定位变化，用重建实例区分“还原输入”和“生成新样本”这两类证据。

表示更有语义、重建更准确、潜在生成更好，是需要分别评价的性质。作者对默认选择的建议属于该工作的主张，不能直接写成全领域已普遍替代VAE的事实。

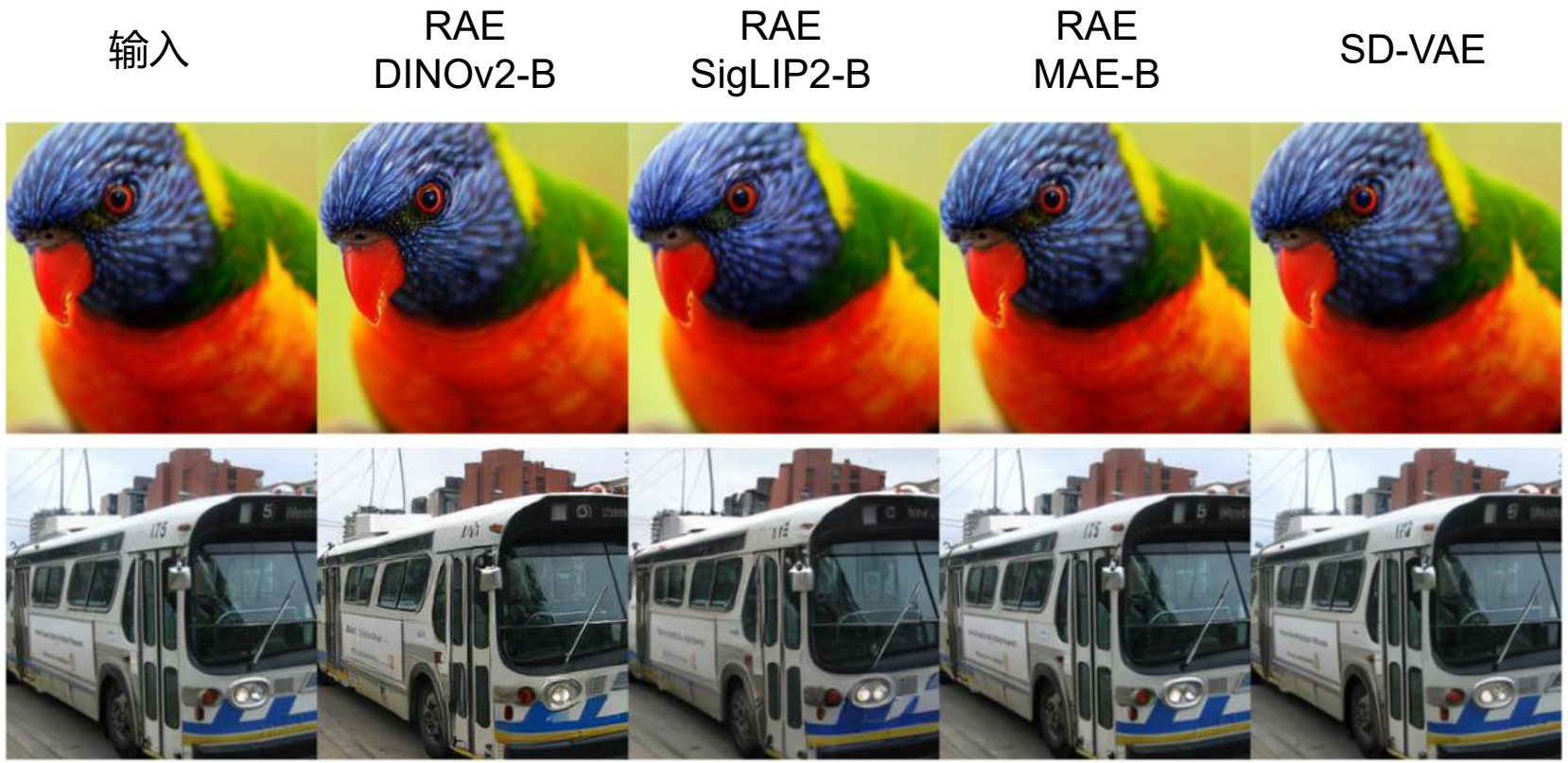
RAE原结构图定位潜在空间的变化



左：SD-VAE；右：DINOv2-B RAE。E编码，D解码。
4为VAE网格通道；768为RAE表示token维度，不能直接作为同一种输入维度比较。
GFLOPs仅属作者256×256设置，不等于端到端延迟；本课未复现。
Zheng等 · 2510.11690v1 · Fig2/§3/AppD · CC BY 4.0

输入与重建终点不构成生成轨迹

按列比较同一输入的重建终点；上方模型名与下方原图逐列对齐。



输入x→潜在表示E(x)→重建D(E(x))；五列为不同模型的终点，不是时间轨迹。新样本仍需独立评价。

Zheng等 · 2510.11690v1 · Fig8 / PDF23 · CC BY 4.0；原图未改，本课未复现。

6 离散状态与掩码恢复

独立选讲：类别转移、吸收掩码与语言案例。

类别转移由概率矩阵规定

如果状态是类别而非连续向量，合法变化需要概率转移来定义。

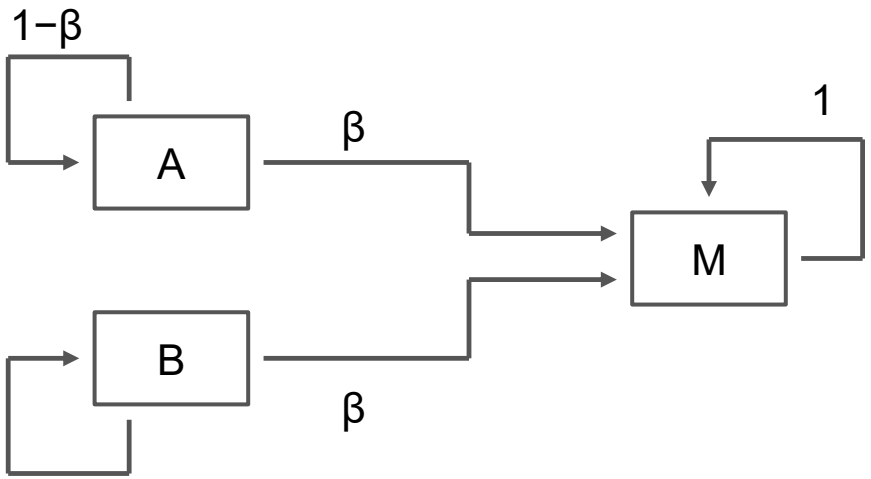
离散状态$x_t \in \{1, \ldots, K\}$由类别而非实数距离定义。D3PM（2021）用转移概率描述逐步破坏过程，本讲采用列向量约定：$M_{ji} = P(x_t = j \mid x_{t-1} = i), \boldsymbol{\pi}_t = \boldsymbol{M}_t \boldsymbol{\pi}_{t-1}$. 每列是给定原类别后的下一状态分布，非负且列和为1。D3PM原文采用行向量与矩阵Q，本讲$\boldsymbol{M} = \boldsymbol{Q}^{\top}$。

类别集合A、B、掩码M的吸收构造为 $\boldsymbol{M}_\beta = \begin{pmatrix} 1-\beta & 0 & 0 \\ 0 & 1-\beta & 0 \\ \beta & \beta & 1 \end{pmatrix}, 0 \le \beta \le 1$. A或B可保持或变成掩码，掩码进入后不再离开；吸收掩码是离散转移的一种选择。

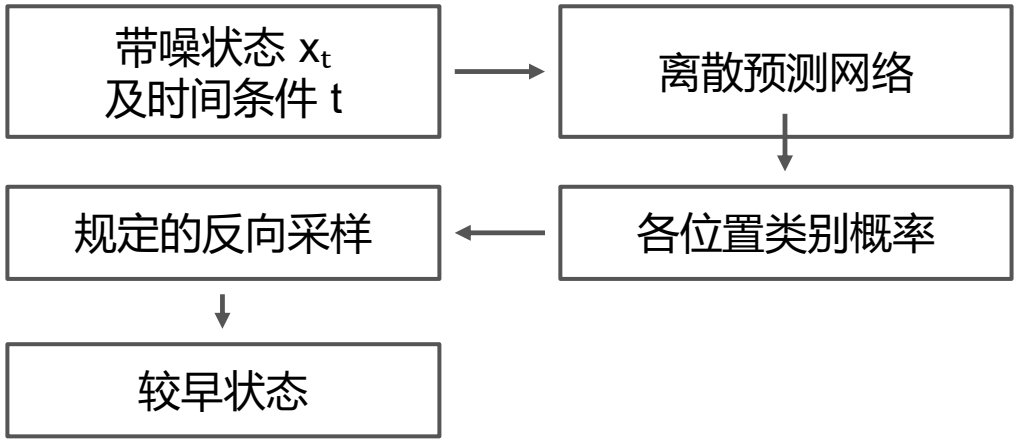
下一页从确定类别A开始，沿相同矩阵计算两次。

吸收状态的路径与反向类别预测

从确定A开始，取$\beta = 0.25$： $\boldsymbol{\pi}_0 = (1,0,0)^{\mathsf{T}}, \boldsymbol{\pi}_1 = (0.75,0,0.25)^{\mathsf{T}}, \boldsymbol{\pi}_2 = (0.5625,0,0.4375)^{\mathsf{T}}$. 同一转移再作用一次，保持A的概率为$0.75^2$，已进入掩码的质量继续保留。



前向吸收；未列边为0



生成不读取真实干净序列

学习反向过程需要预测类别分布，并按已定义模型构造恢复转移。给token编号直接加入高斯，会把任意编号差当成距离，与这个类别转移模型不同。

掩码语言模型改变生成过程，骨干仍可为Transformer

短序列教学示意：“猫 坐 垫”随机掩码为“猫 [M] 垫”。反向模型预测缺失位置的类别；具体填充规则由模型规定。

LLaDA以前向掩码和Transformer反向恢复建模，采用预训练与监督微调。其预测器读取提示与破坏响应，不额外输入t；t用于外部采样规则。

Transformer是网络骨干，生成顺序和训练目标仍需分别说明。下页以原图追踪响应，再解释一个随机重掩码规则，不展开似然下界。

掩码只是离散转移的一种形式。生成速度与质量取决于长度、调用和模型设置，不能从骨干名称推出普遍快于自回归。

LLaDA原图展示全掩码到完整响应

Prompt为固定提示，Response为响应。叉号是掩码，彩块是已填token，颜色不表示具体词义。

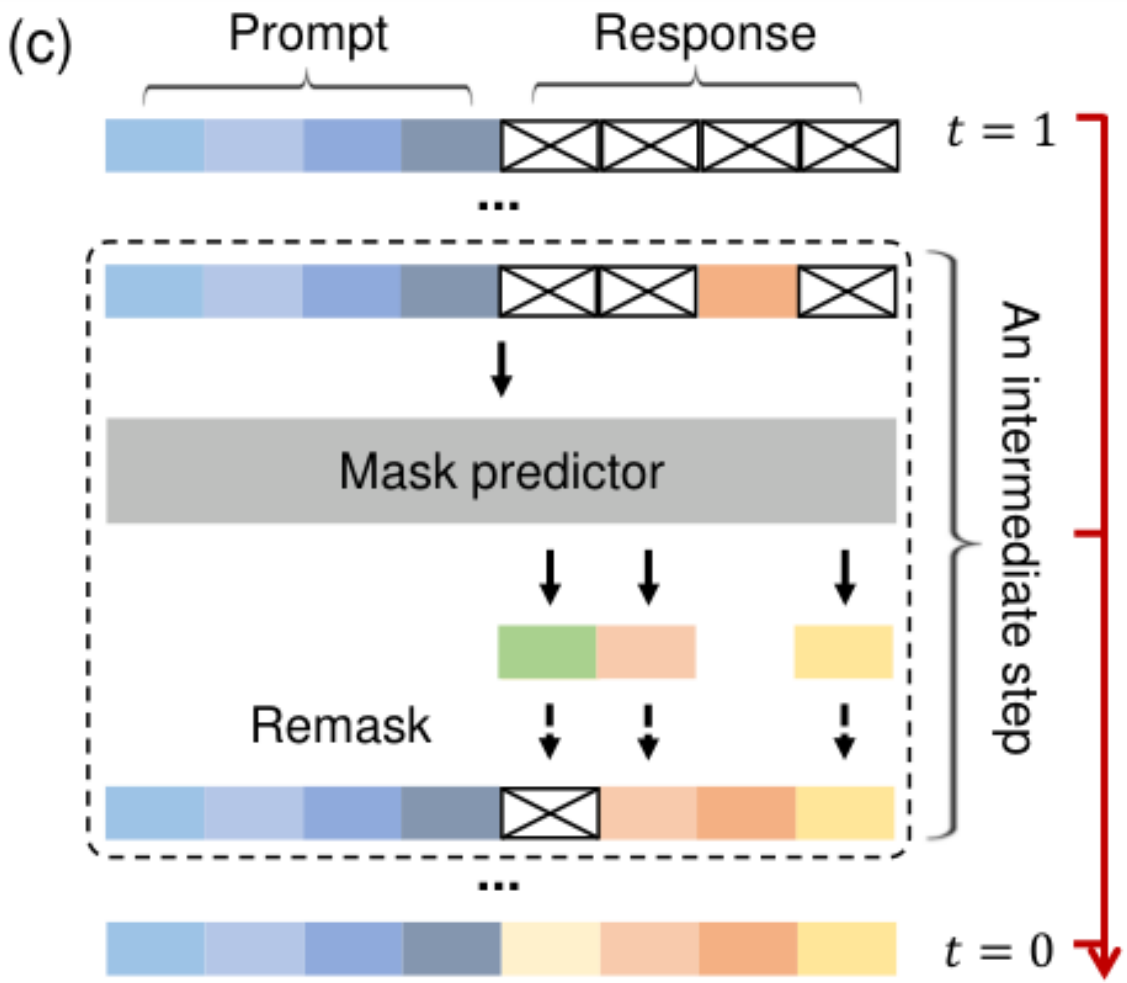
- 开始时t=1，响应全部被掩码
- 中间同时预测缺失位置并重掩码部分候选
- 最后到t=0时，响应中已无掩码

同一响应的概念状态链省略了实际轮数。前向破坏方向为t从0到1。

图中未展示实际文本实验，不能证明质量或速度。

来源：Nie等，2502.09992v1。

图2(c)，PDF第3页。原图教学引用。



随机重掩码规则把预测变为下一状态

据Nie等LLaDA v1 Algorithm 4（PDF第15页）中文译写，采用逐位置记法；这是随机重掩码版本。

固定已训练预测器p_θ、提示p_0、响应长度L和步数N；初始化$r_1 = [M]^L$。

对$t = 1, 1 - 1/N, \ldots, 1/N$，令$s = t - 1/N$：

1. 用$p_\theta(\cdot \mid p_0, r_t)$对所有掩码位置同时取最大概率类别，得到候选$\hat{r}_0$。
2. 对每个位置i：若$r_t^i \neq M$，令$r_s^i = r_t^i$；若仍为掩码，以概率s/t令$r_s^i = M$，否则保留候选$\hat{r}_0^i$。
3. 以r_s作为下一轮状态。到$s = 0$返回无掩码响应r_0。

参数在生成中固定；更新的是响应状态。因最后s/t=0，末轮预测不再重掩码。原文另有低置信度与分块策略，本页只解释Algorithm 4。

生成质量还需真实文本与任务证据；比较成本仍要固定L、N、引导和计数对象。图中全掩码到无掩码只说明过程能够怎样组织。

一次随机重掩码怎样形成下一状态

固定提示、已训练预测器、L=3、N=2。以下是Algorithm 4的一条人工示例路径，候选词不代表真实模型输出。



第一轮：$s/t = 1/2$，每个当前掩码位置独立决定是否继续掩码；示例实现恰好只保留中间的掩码。

末轮：$s/t = 0$，已恢复词保持，剩余掩码用候选替换，返回“猫 坐 垫”。每轮不要求恰好保留固定比例的掩码。

预测器参数不变，更新的是响应状态；比较质量和成本还须固定长度、轮数、引导与任务。

7 视频的时空结构

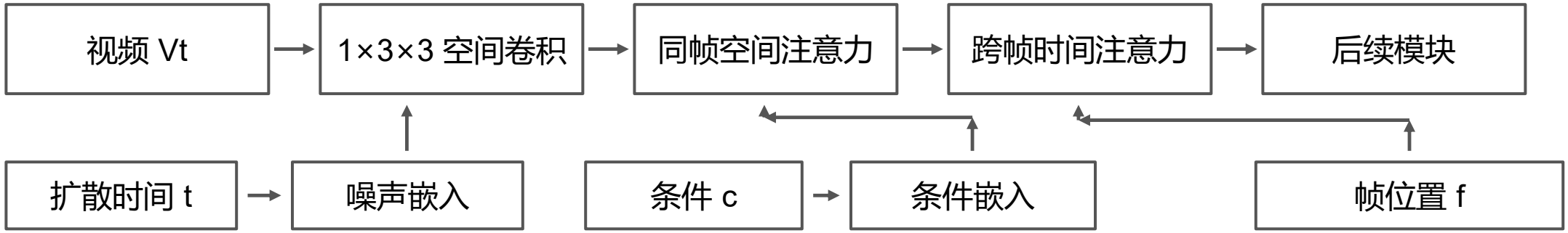
独立选讲：帧时间、扩散时间与跨帧评价。

视频帧维与扩散时间分开组织

从离散序列转到视频，除内容外还增加跨帧关系。

视频状态可以写为 $\boldsymbol{V}_t \in \mathbb{R}^{F\times H\times W\times C}$. F为视频帧数，H、W为空间尺寸，C为通道；下标t为扰动或生成过程的时间，不是视频中的帧编号f。

Video Diffusion Models（2022）的技术基础例在图像U-Net上增加时空分解：空间卷积只处理空间区域，空间注意力在同帧内关联位置，时间注意力在帧维关联信息。原文把3×3图像卷积改为1×3×3，并在空间注意力后插入时间注意力。

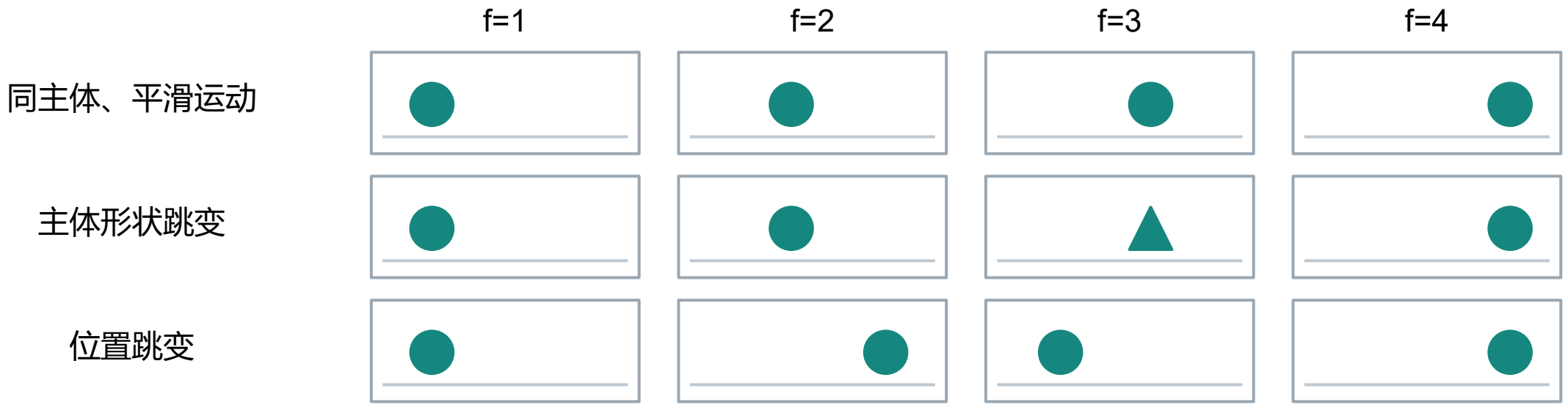


帧时间与扩散时间分别输入；模块按所选方案调制

有了时间模块，仍须用输出帧检查它是否保持主体与运动。

单帧清晰仍可能出现跨帧错误

同一段视频同时有帧内质量与帧间关系。固定四个等间隔帧，下面的几何动画都清晰，跨帧性质却不同。



同一帧间隔、同一坐标尺度；教学构造，不是模型输出

先逐帧看清晰度，再沿同一行追踪主体形状和位置；只有后一条阅读路径才能看到时间异常。

这些是课程构造的可编辑帧示意，不是模型输出或物理仿真。静止画面也可能一致，却没有要求的运动；一致性与动态程度应分别观察。

FVD比较视频分布，VBench拆解评价维度

观察器必须读取要评价的结构：视频片段具有帧间关系，逐帧FID不能替代视频评价。

方法	观察对象与用途	解释边界
FVD	视频片段→固定视频特征→参考 / 生成特征矩比较	受特征、样本和片段协议影响
VBench	主体 / 背景一致、运动平滑、动态程度、成像质量及条件维度	各评估器有误差，不保证全部语义或物理性质

FVD原论文以动作识别I3D特征观察时空信息；它比较整个片段的分布摘要，不是各帧FID的平均。

VBench按维度设置提示与评估方法：先检查主体一致，再看运动与动态，最后核对条件；单个汇总分不能代替分项诊断。

协议同时报告：片段数、帧数、分辨率、帧率、条件、采样器与特征 / 评估器版本。

用评价框架审查一项改进

问题 → 证据 → 协议 → 有限结论：先指出方法改变哪一环，再选择能观察对应主张的证据。

课堂构造：“新模型只需1步，FID更低，因此更快、更懂提示、没有复制。”四个结论都成立吗？

主张	已有证据能支持什么？	还需补什么？
分布更接近	同协议FID支持特征矩更接近	质量 / 覆盖分项与样本
更快	1步只说明更新次数	NFE对象、端到端延迟与条件
更符合提示	FID不读取具体提示约束	对象、属性、关系检查
没有复制	上述两项都不足	训练近邻与来源审计；结论仍有限

可选真实阅读：RAE v1表14，同一DiT^{DH}-XL / DINOv2-B的无引导结果，随机抽类别gFID=1.60，每类固定50张gFID=1.51（均5万张、ImageNet 256×256）。这能归因于网络结构变化吗？