arXiv:2611.29547v1 [cs.CV] 22 Nov 2026

对齐配对项，而非样本：面向更快、更优单步生成器的表征对齐

Mingxiao Li[1] Xuantuo Huang[1,2] Encheng Su[1,3] Aoran Wang[1]Han Zhang[1] Jinouwen Zhang[1] ShixiangTang[1,4]

[1]上海人工智能 gence Laboratory [2]北京大学 [3]中国科学技术大学 [4]牛津大学

https://mingxiao-li.github.io/RePrime/



图 1. **左：RePrime-L 生成的单步 ImageNet** 256×256**样本**（单次网络评估，无引导；FID 为 1.155）。**右：训练计算量（包含教师）与 FID 的关系**（对数坐标轴在 $3\times$ 低于 200 的范围内进行了拉伸）。RePrime 在所有预算和规模下均优于 Priming；RePrime-B 以 $24\times$ 更少的计算量胜过 WF-XL/2。其他结果来自 [26]。

摘要

表征对齐能显著加速扩散训练，但单步生成器引出了扩散训练中不存在的问题：损失应该施加在哪条训练路径上？单步生成器会接收两类输入。受损真实图像有对应的干净图像，可作为实例配对目标。由纯噪声生成的样本则没有对应的真实图像，但最终模型正是在这种输入上运行。我们在 *Priming* 上研究这个问题：*Priming* 通过重建受损图像，再利用最优传输匹配特征分布，从头训练单步生成器。在生成路径上进行对齐——这是自然的选择——会导致所有 *17* 种受测配置的召回率崩溃（$0.45 \to 0.15$–0.25），其中包括五种目标构造方式，以及低至 0.05 的损失权重。相较之下，使用冻结的 *DINOv2*编码器对配对重建路径进行对齐，在我们测试的所有设置中都能带来提升。简短的分析解释了这种不对称性：当目标不包含关于输入的实例信息时，最优的对齐特征在每个类别内都是恒定的，因此损失会将样本拉向类别原型。一项对照实验证实了这一点：若将配对路径对齐到同类别随机样本，而非输入对应的图像，就会失去大部分收益。在生成器内部，配对路径对齐会形成一个接口模块，后训练通过它传递噪声；该模块所承载的图像语义能够预测 *16* 个模型的样本质量。由此得到的配方 *RePrime* 能加速 *ImageNet* $0.45 \to 0.15$ 上的两个训练阶段：它以更少的预热轮次达到 *160* 轮次基线的水平，以更少的后训练更新超越基线的最终质量，并将 *FID* 降至 *1.30*（*B*）和 *1.155*（*L*）。一种联合单阶段变体通过训练原生单步生成器，以 *1.22* 的 *FID* 达到 *1.15* 亿个参数。

1. 引言

扩散模型如今已推动高保真图像和视频合成 [15–17, 27], , 但每个样本都需要数十到数百次网络评估，导致生成速度缓慢，也使误差沿采样轨迹不断累积 [14, 22]。单步生成器通过一次前向传播将噪声映射为图像，从而消除了这项开销。它们可以从扩散教师模型中蒸馏得到 [30, 36, 37] ，也可以从头开始，使用流映射 [6, 8, 9, 21] 和分布式目标函数 [4, 10, 31]进行训练。Priming [26] 显著提升了从头训练的效率：像素空间变换

器首先通过重建严重损坏的图像进行预热，$x_s = (1-s)\,x_0 + s\epsilon$ 使用 $s = 0.8$，随后在纯噪声上进行后续训练，并在冻结的特征空间中通过熵正则最优传输（OT）使生成样本与真实图像相匹配；其在 ImageNet 上达到 FID 1.19，计算量仅为此前单步生成器的十分之一。然而，训练仍然代价高昂，需要 160 个预热轮次，而每次最优传输更新的成本约相当于五次预热更新。扩散训练也面临同样的问题，并找到了有效的解决办法：去噪目标函数学习语义的速度较慢，而 REPA [38] 通过将中间特征与冻结的自监督编码器 [13, 29, 33, 38]对齐，将学习速度提升了一个数量级。这引出了一个自然的问题：表征对齐能否让单步生成器训练得更快、效果更好?

回答这个问题比移植 REPA 更难。在扩散训练中，每个输入都是受损真实图像，因此，对齐目标——编码器对干净图像的表征——始终是明确的。单步生成器的学习输入则有两 种（图 2）。在 配对路径上，它重建一张已知其对应干净图像的受损真实图像；而在生成路径上，它将纯噪声映射为没有 对应真实图像的样本。两条路径都不是显而易见的损失函数适用之处。最终模型在生成路径上运行，而单步目标函数已在冻结特征空间中将样本与真实图像进行比较 [1, 4, 26, 31, 34]。然而，样本没有配对对象，因此必须构造其目标，例如，从随机的同类别图像、最优传输计划与该样本匹配的图像，或样本本身构造；这样的目标可能与 该样本毫无关系，导致同一类别的样本聚集在一起，并削弱分布匹配力图保留的多样性。配对路径有明确的目标，但只存在于训练阶段；与曝光偏差背后的训练—推理间隔类似 [14, 22], , 在受损图像上学到的语义未必能传递到模型用来生成的噪声输入上。此外，由于 OT 阶段已经在冻结特征空间中匹配样本，教师可能只是在重复后训练所见的内容；而且，冻结教师会增加每次更新的计算量，因此任何收益都必须经得起总计算量相同条件下的比较。

为应对上述挑战，我们首先通过可对齐性分析(第3.3节)，探究哪条路径能够支持对齐。如果目标不包含输入的实例信息，那么最优对齐特征在每个类别内都是恒定的，因此损失只能将同一类别的所有样本拉向一个原型；如果输入能够揭示其目标，正如 x_s 仍然能够揭示 x_0所示，最优特征就必须编码实例特定的语义。受此分析启发，RePrime（表征对齐预训练）仅在预训练期间应用配对路径对齐：轻量投影器将中间模块的图块令牌映射到冻结编码器对同一干净图像提取的特征，之后即被丢弃，因此推理过程不变。我们使用跨空间语义教师DINOv2。它不属于 OT 阶段所用的 ConvNeXt 和 MAE 空间，因此能够提供后训练本身无法获得的语义。随后，接口分析通过定位噪声特征与受损输入特征交汇的模块，并测量噪声经由该模块传递了多少图像语义，检验这些语义是否到达生成路径（第4.3节）。最后，RePrime-J 将针对噪声的 OT 更新与对齐重建交错进行，从而得到一个原生单步生成器，无需单独的后训练阶段，同时仍只对齐配对路径。

实验揭示了明显的不对称性。在生成路径上，对齐在我们测试的全部 *17* 种配置中均告失败，涵盖五种目标构造、两个特征空间、两种损失形式，以及低至 0.05 的损失权重：在精确率持平或更高的情况下，召回率从 0.45 降至 0.15–0.25，这正是分析所预测的原型吸引效应。在配对路径上，同类损失在全部 20 项比较中均有帮助；随机同类目标会消除大部分增益，而经由接口传递的语义能够预测 16 个模型的样本质量。在 ImageNet 256×256上，RePrime 加速了两个阶段，以$2\times$ 更少的预热轮次达到 160 轮次基线的水平，并以 $4\times$ 更少的最优传输更新步数超越其最终质量。在全部七种预热预算下以及计算量匹配时，RePrime 均表现更佳（图1右），并将 FID 从 1.37 降至 1.30（B），从 1.17 降至 1.155（L），FD 降低 9–13%$^{5}_{r}$；RePrime-J 使用 1.15 亿个参数达到 1.22 的 FID。

损失本身来自 REPA；我们的贡献在于揭示它应作用于何处、为何如此，以及来自对照组和生成器内部的证据。我们确立了一项放置原则——对齐配对项，而非样本——该原则适用于两阶段和单阶段训练；我们将其转化为 RePrime，这是对 Priming 的即插即用改动，在 [26] 中所比较的方法里取得了最佳单步FID（1.155）；我们还表明，来自最优传输特征空间之外的语义教师至关重要，对齐的容易程度与收益呈负相关，中层目标效果最佳，而且不同于扩散，提前停止对齐并无助益。

2. 相关工作

单步生成。 蒸馏方法将预训练扩散模型压缩为一步或少数几步 [30, 36, 37]。人们通过流映射和平均速度 [6, 8, 9, 21, 41], , 以及直接匹配生成样本与真实样本的分布式目标函数 [4, 7, 10, 19, 31, 34]，重新探索从头开始训练。Priming [26]表明，在固定噪声水平下进行成本低廉的重建阶段，能让随后的分布阶段——在冻结的 ConvNeXt和MAE特征空间中进行熵正则最优传输 [2, 5]——高效得多。与近期的像素空间和频率空间扩散模型 [18, 23],一样，该方法无需学习型分词器。我们要探究，在这一配方中，辅助特征损失应当作用于何处。

用于生成的表征对齐。 REPA [38]将带噪输入的扩散 Transformer 中间特征与冻结编码器提取的干净图像特征进行对齐。后续研究探讨了目标的哪些属性至关重要，发现空间结构十分重要 [29]，何时停止对齐 [33]，使用分词器进行端到端训练 [13]，如何塑造潜在空间本身以提高其可扩散性 [24, 35, 40]，以及无编码器正则化项 [32]。在所有这些方法中，对齐的输入都是受损真实图像，并且有对应的已观测干净图像。单步生成器还会在没有此类对应图像的纯噪声输入上进行训练。这引出了一个在扩散模型中不会出现的问题：应当对齐哪条训练路径？

分布匹配中的表征目标。 SF^2 [1] 为 DMD 添加了预测性表征损失，用于少步视频蒸馏，使用来自共享生成轨迹的视图和一位指数移动平均教师。EASD [28] 将得分和对抗目标分配到不同的噪声区间。我们针对从头开始的单步训练，使用特征空间中的最优传输来回答这一问题：在重建路径上进行实例锚定对齐能够持续带来帮助，而在生成路径上使用无锚定目标则持续以典型性换取覆盖度。语义传输 [31] 表明，对于基于最优传输的单步生成而言，匹配空间至关重要；我们发现，对齐空间也存在互补效应，并且它应当不同于匹配空间。

3. 方法

3.1. 预备知识：Priming

令 $G_\theta(\cdot,\ c)$ 为一个类别条件像素空间 Transformer，它通过一次前向传播 [26] 将 256×256 input 映射为同尺寸图像。Priming 分两个阶段训练 G_θ。

阶段 1（Priming）在固定噪声水平下重建受损真实图像，

$$x_s = (1-s)\,x_0 + s\,\epsilon, \quad \epsilon\sim\mathcal{N}(0,I), \quad s=0.8, \tag{1}$$

通过最小化 $\mathcal{L}_{\text{rec}}(G_\theta(x_s,\ c)\ x_0)$，即自适应归一化的像素误差以及 LPIPS [39] 和 ConvNeXt [20] 感知项。

阶段 2（后训练）输入纯噪声 $z\sim\mathcal{N}(0,\ I)$，并在类别层面逐一将样本分布$X = G_\theta(z,\ c)$与冻结特征空间ϕ_j (ConvNeXt和ResNet-MAE)中的真实图像 Y 进行匹配：

$$\mathcal{L}_{\text{OT}} = \sum_j w_j\Big[D_j\big(\phi_j(X),\phi_j(Y)\big) - D_j\big(\phi_j(X),\text{sg}[\phi_j(X')]\big)\Big], \tag{2}$$

其中，D_j 是通过 Sinkhorn 迭代 [2, 5]计算得到的熵正则最优传输代价，X' 是第二个停止梯度批次样本，w_j 是 [26]的校准权重。在推理时，一次评估 $G_\theta(z,\ c)$即可生成一张图像。

3.2. 两条路径，一个对齐损失

表征对齐 [38] 会在 ℓ块之后，对图像块特征 $h^\ell\in\mathbb{R}^{N\times D}$ 增加一项损失：

$$\mathcal{L}_{\text{align}} = \frac{1}{N}\sum_{p=1}^{N}\Big(1-\cos\big(P_\psi(h_p^\ell),\ T_p\big)\Big), \tag{3}$$

其中，P_ψ 是一个小型可训练投影器，而 $T\in\mathbb{R}^{N\times D'}$是来自冻结编码器的目标映射 E。在 Priming 期间，输入为 x_s，自然目标则是编码器对同一图像的表示，$T = E$（x_0）；我们将其称为配对路径。在后训练期间，输入为 z，且没有与该样本对应的真实图像，因此必须构造 T。我们考虑了所有认为合理的构造方式：以下图像的特征：(i) 同类别的随机真实图像；(ii) 多张同类别图像组成的令牌池；(iii) 最优传输计划 Eq. (2) 分配质量最多的真实图像；(iv) 生成样本本身；以及 (v) 针对相同噪声的指数移动平均教师样本。我们将其称为生成路径。

3.3. 路径为何重要：可对齐性 y

考虑一个目标标记。令 X 表示生成器输入，c 表示其类别，令 $u(X)$表示单位归一化预测，令 T 表示单位归一化目标。对齐风险为 $\mathcal{R}(u) = \mathbb{E}[1-u(X)^\top T]$。

命题 1 (可对齐性)。令 $\mu(X) = \mathbb{E}[T\mid X]$且$m_c = \mathbb{E}[T\mid c]$。则

(i) $\mathcal{R}(u) \geq 1-\mathbb{E}\|\mu(X)\|$，等号成立当且仅当 $u(X) = \mu(X)/\|\mu(X)\|$在 $\mu(X)\neq0$成立的所有位置均如此；*(ii)* 若在给定 c的条件下，T 与 X 相互独立，则最优预测在每个类别内均为常数，$u^\star(X) = m_c/\|m_c\|$；*(iii)* 一般而言，$\mathbb{E}\|\mu(X)\| \geq \mathbb{E}_c\|m_c\|$。间隔 $\mathcal{A} = \mathbb{E}\|\mu(X)\| - \mathbb{E}_c\|m_c\|$ 为零，当且仅当最优预测在每个类别内均为常数。

证明概要。 $\mathcal{R}(u) = 1-\mathbb{E}[u(X)^\top\mu(X)] \geq 1-\mathbb{E}\|\mu(X)\|$ 由柯西–施瓦茨不等式可得，因为 $\|u\| = 1$。在条件独立性下，$\mu(X) = m_c$。第（iii）部分由詹森不等式得出，$\mathbb{E}[\|\mu(X)\|\mid c] \geq \|m_c\|$，当且仅当一个类别内的$\mu(X)$正向共线时取等号。完整证明见补充材料。 □

启示。 该命题虽简单，但其价值在于给出了一个可证伪预测（见第 4.2 节的检验）。超额 $\mathcal{A}$衡量输入中包含多少关于目标的实例信息。在配对路径上，$x_s = 0.2\,x_0 + 0.8\,\epsilon$仍然揭示了 x_0：经过 160 个训练轮次的 Priming 后，生成器对 x_s 的重建在 DINOv2 空间中与其自身的干净图像比与其类别原型更接近（余弦相似度为 0.81 和 0.73；见补充材料）。因此 $\mathcal{A}>0$，而对齐会奖励编码了这张图像语义的特征。在采用随机或池化目标的生成路径上，（ii）完全适用：

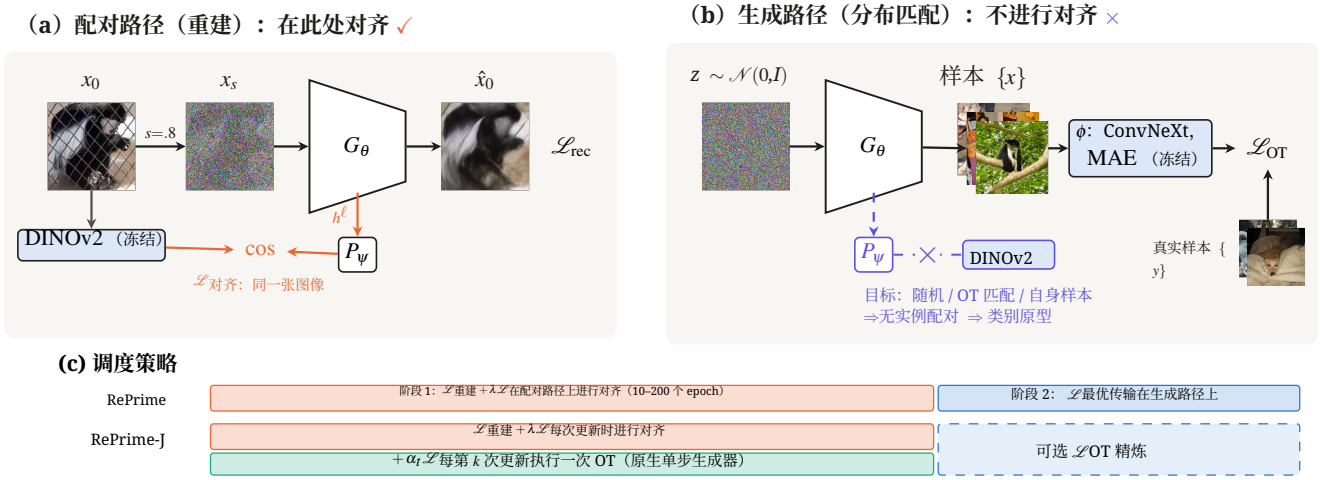


图 2。 **单步生成器应在何处进行对齐。** (a) 在 配对路径上，生成器重建受损真实图像 x_s，并将中间特征 h^ℓ 通过轻量投影器 P_ψ与同一 干净图像 x_0的冻结 DINOv2 特征进行对齐。(b) 在 生成路径上，相同的权重将噪声映射为样本，并在冻结的 ConvNeXt/MAE 特征空间中通过 OT 将样本与真实图像匹配。一个样本没有对应的真实图像，因此我们构造的每个对齐目标都缺乏实例对应关系，并导致覆盖度坍缩（第 4.2节）。(c) RePrime 仅在 Priming 期间进行对齐；RePrime-J 联合训练两条路径。

最优解是一个类别恒定方向，梯度会将每个类别为 c 的噪声输入的特征都拉向同一个原型。OT 匹配目标仅通过在 ConvNeXt/MAE 空间中计算出的传输方案依赖于样本；在 DINOv2 空间中，匹配图像与样本的接近程度仅略高于同类随机样本（余弦相似度为 0.221 和 0.213；见补充材料）。自目标和 EMA 目标会随生成器变化，因而不属于该命题的适用范围，但也呈现出同样的症状：精确率更高，召回率却发生坍塌（见第 4.2 节）。

3.4. RePrime：对齐配对项，而非样本

RePrime 仅在配对路径上应用公式（3）：

$$\mathcal{L}_{\text{Stage 1}} = \mathcal{L}_{\text{rec}}\big(G_\theta(x_s, c), x_0\big) + \lambda\, \mathcal{L}_{\text{align}}\big(h^\ell(x_s, c),\, E(x_0)\big), \tag{4}$$

并保持阶段 2（公式（2））不变。 E 采用 DINOv2-L/14 [25]，，应用于 x_0 并调整大小为 224^2。其 16×16 图块令牌与生成器的图块网格逐一对齐（不包括类别令牌）。 P_ψ 是一个两层 GELU 多层感知机（宽度为 1024），使用 AdamW 训练（学习率为 10^{-4}），并在 Priming 后丢弃，因此推理过程不变。我们使用$\lambda = 0.5$，在整个 Priming 期间保持对齐开启，并对 B 的 16 个区块中的区块 $\ell = 5$ 进行对齐（区块 8 更好；第 4.6节），对 L 的 32 个区块中的区块 10 进行对齐。

RePrime-J：联合预热。 两阶段方案在一条路径上学习语义，之后才教会另一条路径使用这些语义。如果正如我们在第4.3节所示，后训练主要是将噪声引导通过预热期间形成的特征，那么两条路径也可以同时训练，使噪声在这些特征形成的同时学会使用它们。RePrime-J 交错进行两个目标函数；在外层更新 t时，

$$\mathcal{L}_t = \mathcal{L}_{\text{Stage 1}} + \alpha_t \mathbb{1}[t \bmod k = 0]\, \mathcal{L}_{\text{OT}}, \tag{5}$$

将 $k = 2$ 和 α_t 在前 25k 次更新中从 0 线性增加到 1。两个损失使用独立的前向传播和一次优化器步。尽管生成路径现在贯穿始终地进行训练，对齐仍仅作用于配对路径：无论调度如何，命题 1 均适用，而生成路径仅通过共享权重获得教师语义。由此得到的 原生 模型无需后训练，但可以通过阶段 2 进行优化。

计算成本。 教师模型为每张训练图像额外增加一次 DINOv2-L 前向传播（162 GFLOPs），即， +43% 浮点运算用于每次 B预热更新，以及 +19% 用于 L；投影器的计算量可以忽略不计。由于在低预算下阶段 2 占据主要计算量，因此这相当于 B 规模总训练计算量的 +4–29% ，以及 L 规模总训练计算量的 +14%（表 1）。一次 OT 更新的成本约为4.9× 一次预热更新，因此 RePrime-J 的 $k = 2$ 成本约为3.9×一次 Priming 更新。按照 [26]，我们将所有成本均以 EFLOPs 报告。

4. 实验

4.1. 设置

我们遵循 Priming [26]在类条件 ImageNet256×256 [3]上的方案，使用其像素空间 B（1.15 亿参数）和 L（4.05 亿参数）生成器，以及对应的配方：160 个预热训练轮次（200,160 次更新），随后按照公式（2）进行 25,000 次后训练更新（超参数见补充材料）。RePrime 与基线方法的唯一区别在于公式（4）中的损失。基线方法对发布的 Priming-B/L 初始化进行后训练，从而复现报告的 FID 分数；预算扫描则重新训练官方方案。

指标。 我们通过一次网络评估生成 50k 个样本，且不使用引导。我们按照 [26]的官方协议报告 FID [11] 和 FD_r^5。FD_r^5 对 Inception 中归一化的 Fr´echet 距离取平均值

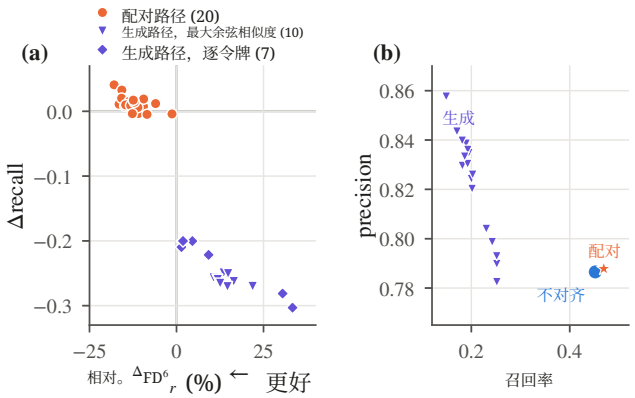


图3. **对齐位置决定了对齐的影响方向**。(a) 与相应的无对齐基线相比，FD^6_r 和召回率的变化：20 项配对路径比较和 17 种生成路径配置。(b) 在相同的 2.5k 更新步数方案下的精确率–召回率。



图4。**原型吸引**（类别 207，每行使用相同的 8 个噪声向量，2500次训练后更新）。第1–4行：Priming-B 初始化，分别采用无对齐、与 OT 匹配目标的生成路径对齐、与同类随机样本目标的生成路径对齐，以及与 EMA 教师目标的生成路径对齐。第5行：RePrime 初始化。OT 匹配行和随机行几乎相同，都是居中、近景的狗，背景也更简单（例如，第 2、5、8列）。补充材料中展示了更多类别。

在 Inception、DINOv2、ViT-MAE、SigLIP 和 CLIP 空间中的距离 [34]；FD^6_r 还纳入了 ConvNeXt，并附带精确率和召回率 [12]，这些数据来自训练时评估器。主要结果取三个采样随机种子的平均值；FD^6_r、精确率和召回率使用 EMA 权重。

4.2. 对齐位置：配对路径与生成路径

方案。从发布的 Priming-B 初始化开始，我们使用公式 (2) 对区块 8 进行后训练，并加入对齐损失；在 250 次更新后启用该损失。我们在 DINOv2 和 OT ConvNeXt 空间中测试第 3.2 节的五个目标，采用标记上的最大余弦相似度（10 种配置），以及对随机目标和 OT 匹配目标采用逐令牌余弦相似度，$\lambda \in \{1, 0.25, 0.1, 0.05\}$，并使用在 750 次更新后结束的窗口（7 种配置）。如果生成路径对齐只是调优不佳，那么其中一些配置应该会有所帮助；如果命题1 正确，那么所有配置都会将样本拉向类别原型，从而提高精确率并降低召回率。作为相同方案下的配对路径对应项，我们使用普通公式 (2) 对 RePrime-B 初始化进行后训练。

生成路径对齐以典型性换取覆盖度。全部 17 种配置都使召回率从 0.452 降至 0.149–0.252，并使 FD^6_r (14.71→14.92–19.62) 恶化；与此同时，其中 16 种配置的精确率仍保持在基线水平或更高（图 3b）。精确率上升而召回率下降，意味着样本集中于该类别的典型模式，而不是各个样本本身变得更差：这正是命题 1所预测的原型吸引效应。该失败并非剂量问题：在 λ=0.05时，召回率仍降至 0.252。它也不是暂时性的：在 750 个更新步后关闭对齐并继续进行 1750 个更新步的纯 OT，召回率仍为 0.242；而在 λ=0.05 对齐后再进行 2.5k 个纯 OT 更新，召回率也仅从 0.252 升至 0.290，这表明 OT 只能挽回少量损失。此外，这也不是目标空间选择不当：在 OT ConvNeXt 空间中，投影器几乎完美地匹配了其目标（余弦相似度为 0.998），但召回率仍降至 0.18–0.20，因此，一个拟合良好的目标如果没有关于样本的信息，也无济于事。图 4展示了这一影响；OT 匹配目标表现得像随机目标，因为从 DINOv2 的角度看，该计划在类别内部几乎是随机地选择配对对象（第 3.3节）。

配对路径对齐提升保真度和覆盖度。在相同方案下，RePrime 初始化使FD^6_r 从 14.71 降至 12.89，并使召回率从 0.452 升至 0.469，而精确率保持不变。在本文全部 20 项配对路径比较中（预算、尺度、深度、教师、后训练时长、联合训练），FD^6_r 改善了 1–18%，召回率下降幅度从未超过 0.005（图 3a）。因此，同一种损失在一条路径上会压缩覆盖度，在另一条路径上却能改善覆盖度；结果取决于损失的施加位置，而非损失本身。**实例对应关系带来了增益**。上述比较同时改变了路径和目标，因此我们仅在配对路径上检验命题1：使用破坏对应关系的目标（同类别随机图像，或令自身图像的标记发生置换）进行 10 个训练轮次的 Priming，然后进行 5000 次后训练更新。如果对齐的作用只是提供通用正则化，或注入类别级语义，那么随机目标应该能恢复大部分增益。但事实并非如此：FD^6_r 在无对齐时为 22.10，在两次随机目标运行中分别为 21.62 和 21.47，使用置换后的标记时为 19.94，使用自身图像时为 18.10，也就是说，随机目标只能恢复 12–16% 的增益，置换后的标记则能恢复 54%。召回率也呈现相同趋势（从 0.486 升至 0.514）。排名与目标中的实例信息相吻合：随机图像不携带实例信息；置换后的标记保留图像内容，但丢失其布局；自身图像则两者兼具。因此，只有当目标描述特定输入时，对齐才有帮助。

4.3 配对路径对齐为何有效

对齐仅作用于受损真实图像，但在从噪声生成图像时却能带来增益。由于后训练仅将预热权重移动了 2–4%（补充材料），因此它所利用的因素必然存在于 Priming 之后。于是，我们开始探究噪声

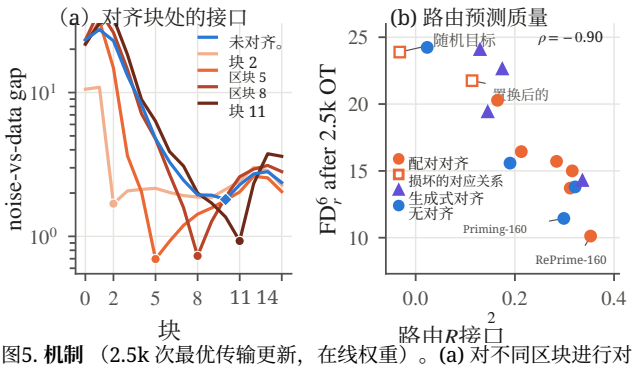


图5. **机制**（2.5k 次最优传输更新，在线权重）。(a) 对不同区块进行对齐的 40 训练轮次 RePrime 模型中，噪声与数据特征差距随深度的变化（圆点：对齐块）。(b) 16 个模型在接口处的语义路由与 FD_r^6 的对比。

和真实样本在生成器内部何处相遇，以及有什么信息经过该位置：使用相同的标签，将 16 个类别的 1,024 张图像作为数据输入 x_s 或作为噪声 z 输入，并比较均值池化块特征。

对齐会创建一个接口。 噪声与数据差距是这两组特征之间的 Fr´echet 距离，以数据协方差的迹为基准。经过 2.5k 次最优传输更新后，在四种对齐深度下，差距都恰好在对齐块处最小（图 5a），且比无对齐时的最小值小 2。6×（0.70 对 1.81）。由于最小值会随对齐块移动，因此接口是由对齐创建的，而非架构的固定属性。在后训练之前，接口就已存在（训练 160 个轮次后，区块 5 处的差距为 4.5 对 28.4），因此后训练是找到了它，而非构建了它。

重要的是经由它路由的语义。 仅凭较小的间隔并不能解释质量：生成路径的对齐也会缩小间隔（0.47–0.61），这主要是通过将同一类别的噪声特征挤到一起实现的（类内扩散为 0.13，而非 0.41）；在 16 个模型中，间隔与F D_r^6（斯皮尔曼 ρ=0.00）不相关。因此，我们测量语义路由：在接口处用数据模式特征拟合一个岭回归探针，以预测干净图像的 DINOv2 特征；随后将其应用于噪声模式特征，并根据生成图像的 DINOv2 特征评分。较高的 R^2 意味着噪声像真实输入一样使用该接口。路由与FD_r^6 在全部 16 个模型中均有相关性（ρ=−0.90；图5b）。对应关系对照将这两个量区分开来：随机目标也会在区块 5 形成接口（间隔为 1.57），但其特征对留出集 CLIP 嵌入的预测效果并不优于未进行对齐时（R^2 00.361 对 0.360；自身图像为 0.413），而噪声也不会通过该接口路由任何语义（−0.03 对比 0.17）。这完整阐明了放置原则：实例配对目标会将特定图像的语义写入接口，而后训练则将噪声经由该接口路由；非配对目标则会创建一个没有有用信息可供路由的接口。

方法	Ep.	EFLOPs	FID↓	FD_r^5 ↓	精确率↑	召回率↑
报告值为 *[26]*（*1 NFE*）						
pMF-H [21]	–	1968.5	2.15	8.14	–	–
iMF-XL/2 [9]	–	3201.1	1.74	9.75	–	–
Drifting-L [4]	1280	1476.4	1.61	12.59	.815	.602
WF-XL/2 [10]	–	3744.4	1.33	11.07	.762	.674
Priming-B [26]	160	124.2	1.38	11.23	.783	.624
Priming-L [26]	160	239.0	1.19	7.10	.781	.649
B，*1.15*亿个参数（我们的运行结果，流程完全相同）						
Priming-B 160	124.2	1.373	11.27		.784	.620
Priming-B	200	143.4	1.353	10.91	.784	.625
RePrime-B	120	129.8		10.76	.784	.627
RePrime-B	160	157.3		10.20	.781	.639
RePrime-B	200	184.8		9.66	.784	.631
RePrime-B，区块 8	160	157.3		9.82	.781	.641
RePrime-B，区块 8	200	184.8		9.66	.782	.635
RePrime-J-B*	160	346.8		10.60	.779	.640
L，*4.05*亿参数（我们的运行，采用相同流程）						
Priming-L	160	239.0	1.174	7.41	.779	.648
RePrime-L	160	272.1		**6.43**	.781	.644

表 1。**ImageNet** 256×256，**一次网络评估，无引导**。EFLOPs 统计两个阶段的计算量；对于 RePrime，还包括冻结教师。“Ep.”：预热轮次；所有模型均经过 25k 更新步数的后训练。FID/FD_r^5 为三个采样随机种子的平均值（120 训练轮次这一行使用一个随机种子；标准差 ≤0.022 FID，≤0.035 FD_r^5）。除非另有说明，RePrime-B 对齐区块 5。 *联合训练（第 4.7 节），进行 200k 次联合更新，随后进行 25k 步精炼更新。

4.4. 主要结果

表 1 在相同流程下比较了 RePrime 与 Priming。在 B 规模和 160 个训练轮次下，RePrime 将 FID 从 1.373 降至 1.303，并将 FD_r^5 降低了 9.4%（11.27→10.20），同时在精确率不变的情况下将召回率从 0.620 提高至 0.639；与第 4.2节配对路径上的情况一样，对齐改善了样本质量，而不会使样本过度集中。RePrime-L 达到 1.155 的 FID，是所报告方法中最佳的单步 FID，优于 [26],下方的 Priming-L（1.19）和 WF-XL/2（1.33）；后者使用了14× 更多计算量；即使是 RePrime-B，也以 24× 更少的计算量超过了 WF-XL/2（图 1）。L 规模下的 FID 增益（0.019）处于随机种子变化范围内，但语义指标表明，该收益不会随规模扩大而缩小：FD_r^5 降低了 13.3%（7.41→6.43），而相对 FD_r^6 增益从 −9.5%（B 规模）提高至 −12.7%（L 规模）。

提升幅度超过运行间差异。 每种配方进行的第二次 160 训练轮次 Priming 运行与第一次结果相差不超过 0.21 FD_r^6 和 0.006 FID，而不同配方之间的差异为 0.9– 1.2 FD_r^6 和 0.07 FID（补充材料），是训练噪声的四到六倍。

这些提升并非教师模型伪影。 DINOv2 是我们的教师，也是五个 FD_r^5 空间之一，因此提升可能只是反映了与教师的一致性。去掉 DINOv2 项后，B 处的 FD_r^5 仍下降了 8.6%（12.46→11.38），L 处下降了 12.7%（8.19→7.14）；在 L 处，所有非教师空间的表现都有提升，在 B 处则除 CLIP 外均有提升。因此，RePrime 在从未见过的编码器评估下也能改善样本；对于相同的噪声，它能生成更连贯的形状和布局，尤其是在预算较低时

（图6）。

4.5. 更快的训练

更少的预热轮次。 图 7(a,b) 将预热预算从 10 个训练轮次变到 200 个训练轮次，每种设置后都进行 25k 次训练后更新。RePrime 在全部七种预算下的 FID、FD^5_r、FD^6_r 和召回率均更优。预热 80 个训练轮次后，其表现可与 160 个训练轮次的基线持平（FD^5_r 11.24 对比 11.27；FD^6_r 9.43 对比 9.51），预热时间减半；预热 120 个训练轮次后，其所有指标均超过 200 个训练轮次的基线。相对提升在小预算时最大（$-16\%FD^5_r$，预热 10 个训练轮次时），此时基线学到的语义最少，但提升在 200 个训练轮次时依然存在。因此，对齐不仅能更快达到相同的模型水平，也能改善较长时间预热最终得到的模型。

在计算量匹配的比较中，提速优势依然存在。教师使每次 B预热更新的成本增加了 43%，因此每次更新取得更快进展，并不一定意味着训练更快。但本例确实如此（图7c）：训练 80 个训练轮次的 RePrime-B 比训练 120 个训练轮次的 Priming-B 少用 2.5% 计算量（102.3 对 105.0 EFLOPs），且 FID 更优（1.404 对 1.415），FD^5_r（11.24 对 11.82），以及FD^6_r（9.43 对 9.97）。

减少后训练更新步数。如果后训练是在 Priming 阶段形成的接口基础上进行（第4.3节），那么更好的初始化也应缩短耗费较高的最优传输阶段，事实确实如此（图7d）。从训练 160 个训练轮次得到的初始化出发，RePrime 经过 3.75k 次更新后，FD^6_r 便达到基线的最终值（9.51），并在 6.25k 次更新后超越基线，$4\times$ 所需更新步数更少。在 40 个训练轮次时，RePrime 经过 2.5k 次更新后，效果就与基线经过 22.5k 次更新后的结果相当（在线FD^6_r 13.71），更新步数 $9\times$ 有所减少。一次 OT 更新的成本相当于$\sim$4.9 次 Priming 更新，因此这是最大的实际节省。后训练时间更长也不会抹去这一优势，它并非只是抢先起跑：经过 50k 次更新后，差距从 0.90 扩大到 1.11 FD^6_r（8.04 对 9.14），这表明后训练持续利用而非取代 Priming 阶段学到的内容。

4.6. 对齐什么、在哪里对齐以及何时对齐

教师：它必须补足 OT 阶段所欠缺的内容。 表。2比较了 DINOv2 与 OT损失已经使用的两个特征空间。在 40 个训练轮次时，DINOv2 使 FD^6_r提高了 16.5%，ResNet-MAE 提高了 6.0%，而 ConvNeXt 仅提高了 1.3%；在完整预算下，MAE 和 ConvNeXt 对齐甚至略有损害（9.65 和 9.95，而非 9.51），而 DINOv2 则带来改善（8.61）。分空间分析（补充材料）解释了其中原因。ConvNeXt 对齐使 ConvNeXt 距离改善了 17%，但使 CLIP 恶化了 8%，即，它强化了后训练已经在优化的内容。相反，DINOv2 对齐使 CLIP 和 SigLIP 提高了 18–20%，而训练阶段和教师都没有使用这两者。对齐的容易程度与收益成反比（最终余弦相似度分别为 0.97、0.80 和 0.55）：对于投影器几乎能够完美匹配的目标，它对生成器特征施加的压力很小。一个有用的教师必须具备语义性，并且位于 OT空间之外；这样

	设置	对齐余弦	FD^6_r ↓	召回率↑
*40*轮预热，*2500*次训练后更新（指数移动平均）				
teacher	无对齐 – 17.71			.416
	DINOv2-L（跨空间）	0.55	**14.79**	.427
	ResNet-MAE（OT空间）	0.80	16.65	.428
	ConvNeXt（OT空间）	0.97	17.49	.412
window	Priming 的前 25%	–	15.89	.422
	Priming 的前 50%	–	15.52	.426
	完整 Priming	–	**14.79**	.427
depth	块 2 / 16	–	15.78	.413
	区块 5 / 16	–	14.79	.427
	区块 8 / 16	–	**14.54**	**.457**
	块 11 / 16	–	14.93	.449
160/200 个训练轮次的 *Priming*，*25k* 后训练更新（指数移动平均）				
teacher	无对齐 – 9.51			.620
	DINOv2-L	–	**8.61**	**.639**
	ResNet-MAE	–	9.65	.622
	ConvNeXt	–	9.95	.619
depth	区块 5，160 / 200 周期	–	8.61/ 8.19	.639/ .631
	区块 8，160 / 200 个周期。	–	**8.28**/ **8.15**	.641 / .635

表2。**RePrime-B 消融实验。**“对齐余弦”：Priming 结束时的训练对齐余弦。

		原生		+25k 最优传输更新		
配方	EFLOPs	FID	FD^5_r	FID	FD^5_r	FD^6_r
Priming-B（两阶段）	124.2	–	–	1.373	11.27	9.51
RePrime-B（两阶段）	157.3	–	–	1.303	10.20	8.61
Joint-B	313.6	**2.06**	11.00	1.265	11.52	9.72
RePrime-J-B	346.8					

表3. **单阶段训练**（B，160 个训练轮次，即，200k 次联合更新）。“原生”：无后训练（指数移动平均，单个随机种子）。EFLOPs 包含精炼和教师。

对齐就能补充后训练，而不是重复后训练。

窗口：不要过早停止。 在扩散中，提前停止 REPA 有帮助 [33]。在 Priming 中，收益则随窗口单调增长（FD^6_r 15.89、15.52 和 14.79，分别对应 25%、50% 和完整窗口）。一种可能的解释是，Priming 在单一高噪声水平下训练，此时语义目标始终具有信息量。

深度：网络中部。 质量随深度呈倒U形，在 40 个训练轮次时，16 个块中的区块 8 在 FD^6_r 和召回率两项上均表现最佳。一种可能的解释是，早期块对语义而言过浅，而后期块负责像素解码。这一排名在完整预算下仍然成立：区块 8 在 160 个训练轮次时达到 8.28，在 200 个训练轮次时达到 8.15，是使用 25k 次后训练更新时最佳的 FD^6_r；因此，我们默认使用区块 5，低估了可实现的增益。

4.7. 联合单阶段训练

为何要同时训练两条路径。在两阶段方案中，Priming 将语义写入接口，而后训练则教会噪声使用该接口（第4.3节）。由于后训练只会使权重变化 2–4%，生成路径主要只需找到该接口，没有任何理由要求它必须等到



图 6. **相同噪声，不同初始化**。 Priming（上）和 RePrime（下）在训练 40、80 和 160 个训练轮次后的未筛选样本 f Priming 后进行 25k 次后续训练 训练更新步数（B），以及 L 规模下（右）。每列使用相同的类别和噪声。

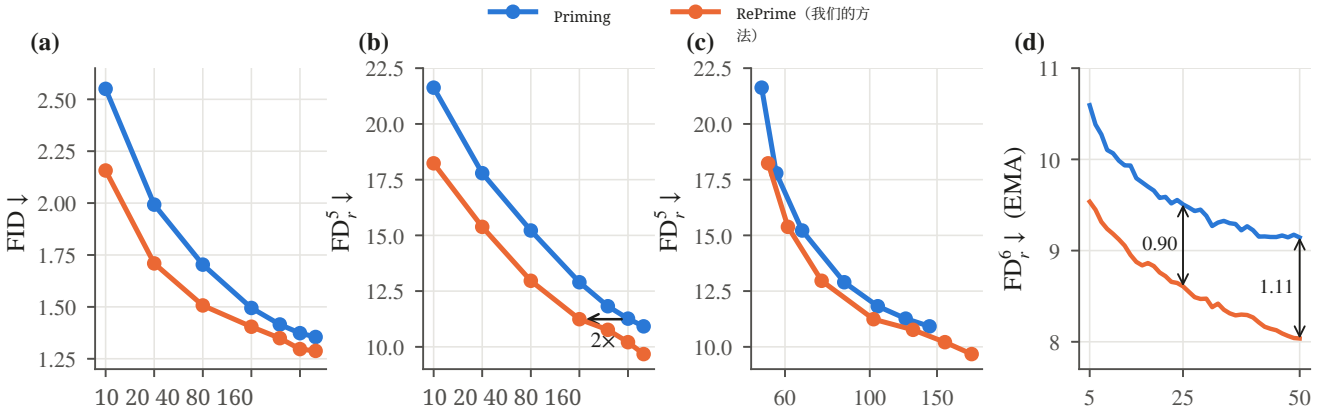


图 7. **RePrime 加速了两个阶段**。 (a,b) 后训练更新 25k 步后，FID 和 FD_r^5 随预热轮次的变化。(c) FD_r^5 随总训练计算量的变化，包含教师的计算量。(d) 以 160 个训练轮次初始化的后训练曲线。

Priming 结束：噪声可以在接口形成的同时学会使用它，从而得到一个原生单步生成器。联合训练（公式（5））检验了这一点，也检验放置原则是否取决于调度。由于最优传输现在从一开始就塑造生成路径，它可能会自行提供语义，使对齐变得多余；但命题1 则预测，配对路径对齐仍然有帮助。

当两条路径共同训练时，对齐仍然至关重要。 在一个训练 40 个轮次的 2×2 析因设计中，每种设置之后都进行简短的 1.25k 步最优传输探测。对齐将顺序方案的 FD_r^6 从 18.11 提升至 14.61，将联合方案的 FD_r^6 从 15.11 提升至 12.44（见补充材料）。仅联合训练也有帮助（18.11→15.11），而两者结合效果最佳。这些效应并非冗余：对齐带来的提升仅从 3.50 缩小到 2.67，因此早期最优传输并不能提供教师语义。配对路径通过对齐决定接口编码什么；生成路径则决定噪声如何到达接口。

原生单步生成。在完整预算下（表3），原生 RePrime-J 在 20 万次联合更新后达到 FD_r^5 9.75，240k 次后达到 9.33，优于所有两阶段 B 模型（最佳为 9.66），且无需任何后训练。原生 RePrime-J 在从 5 万到 24 万次更新的每个检查点上，FD_r^5 都优于原生联合训练（5 万次时为 13.96 对 16.71）；约 11 万次更新后，它便超过两阶段 Priming-B（11.27），而不进行对齐的联合训练则需要约 19 万次。原生 FID 呈现相反趋势（20 万次时为 2.47 对 2.06），但差距持续缩小，从 5 万次时的 1.16 降至 24 万次时的 0.17。我们认为，这是低层统计量收敛较慢所致，而 Inception 对这类统计量较为敏感；与此同时，由 FD_r^5 衡量的语义表现持续改善。

精炼与成本。 使用 25k 次最优传输更新进行精炼，可获得最佳的 B 规模 FID（1.224，而未进行对齐时为 1.265），并使 FD_r^6 降低 8.5%（9.72→8.90）；这一方向及幅度与两阶段方案中的结果（9.51→8.61）相同且相近。精炼使 FD_r^5 高于原生模型（9.75→10.60），因此，在 ConvNeXt/MAE 空间中进行最优传输，会以部分留出语义为代价换取保真度。在 $2.8\times$ Priming-B 的成本以及 $2.2\times$ RePrime-B 的成本下，RePrime-J 并不是成本更低的方案；在固定预算下，两阶段 RePrime 是更好的选择。其作用在于说明，即使生成路径始终处于激活状态，对齐配对项，而非样本 这一点仍然成立：教师语义通过共享权重传递至噪声，而不是通过针对样本的目标传递。

5. 讨论与结论

局限性。 我们的研究涵盖了类别条件的 ImageNet、256^2，Priming架构和 OT后训练；文本到图像的生成以及其他目标函数仍有待测试。大多数配置只训练了一次，但重复运行（第4.4节）以及七种预算下呈现的一致趋势表明，结论具有稳健性。我们的生成路径目标函数都缺乏实例对应关系；能够自行建立对应关系的目标函数，例如，如SF 中的轨迹一致性视图[2] [1]，，仍有待研究。最后，教师模型会增加计算量；使用缓存特征可以降低这一开销。

结论。 表征对齐只能帮助单步生成器的两条路径之一。在配对路径上，

它将图像语义写入一个接口，后训练通过该接口传递噪声，使预热轮次减半、最优传输更新减少至四分之一；在生成路径上，我们尝试的每个变体都以覆盖度换取典型性。我们希望对齐配对项，而非样本这一理念能够指导表征学习与分布匹配的结合方式。

大语言模型使用声明

研究中使用了大语言模型助手来核对数字、编写绘图代码和编辑文本。作者设计并开展了所有实验，核实了所有数字，并对内容承担全部责任。

参考文献

[1] Junyi Chen、Haiyu Zhang、Zhoujie Fu、Tengfei Wang 和 Chunchao Guo。《超越分布匹配：面向少步视频生成的自监督表征强制》。arXiv 预印本 arXiv:2610.35594，2026年。2、3、8[2] Marco Cuturi。《Sinkhorn距离：最优传输距离的高速计算》。载于《神经信息处理系统进展》，第2292–2300页，2013年。2、3[3] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。《ImageNet：大型层级图像数据库》。载于《2009年IEEE计算机视觉与模式识别会议》，第248–255页，2009年。4[4] Mingyang Deng、He Li、Tianhong Li、Yilun Du 和 Kaiming He。《通过漂移进行生成建模》。arXiv 预印本 arXiv:2602.04770，2026年。1、2、6[5] Jean Feydy、Thibault S´ejourn´e、Franc。ois-Xavier Vialard、Shun-ichi Amari、Alain Trouv´e 和 Gabriel Peyr´e。《使用Sinkhorn散度在最优传输与最大均值差异之间插值》。载于《第二十二届国际人工智能与统计会议论文集》，第2681–2690页。PMLR，2019年。2、3[6] Kevin Frans、Danijar Hafner、Sergey Levine 和 Pieter Abbeel。《通过快捷模型实现一步扩散》。载于《国际学习表征会议》，2025年。1、2[7] Aude Genevay、Gabriel Peyr´e 和 Marco Cuturi。《使用Sinkhorn散度学习生成模型》。载于《第二十一届国际人工智能与统计会议论文集》，2018年。2[8] Zhengyang Geng、Mingyang Deng、Xingjian Bai、J. Zico Kolter 和 Kaiming He。《用于一步生成建模的均值流》。arXiv 预印本 arXiv:2505.13447，2025年。1、2[9] Zhengyang Geng、Yiyang Lu、Zongze Wu、Eli Shechtman、J. Zico Kolter 和 Kaiming He。《改进的均值流：快速生成模型面临的挑战》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2026年。1、2、6[10] Jiaqi Han、Puheng Li、Qiushan Guo、Renyuan Xu、Stefano Ermon 和 Emmanuel J. Cand`es。《通过Wasserstein梯度流进行一步生成建模》。arXiv 预印本 arXiv:2605.11755，2026年。1、2、6

[11] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bern- hard Nessler 和 Sepp Hochreiter。通过双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡。载于《神经信息处理系统进展》，2017年。4[12] Tuomas Kynk¨a¨anniemi、Tero Karras、Samuli Laine、Jaakko Lehtinen 和 Timo Aila。用于评估生成模型的改进精确率和召回率指标。载于《神经信息处理系统进展》，2019。5[13] Xingjian Leng、Jaskirat Singh、Yunzhong Hou、Zhenchang Xing、Saining Xie 和 Liang Zheng。REPA-E：借助潜在扩散 Transformer 解锁变分自编码器的端到端调优。载于《IEEE/CVF国际计算机视觉会议论文集》，2025年。2, 3[14] Mingxiao Li、Tingyu Qu、Ruicong Yao、Wei Sun 和 Marie-Francine Moens。通过采用偏移时间步采样缓解扩散模型中的曝光偏差。载于《国际表征学习会议》，2024年。1, 2[15] Mingxiao Li、Bo Wan、Marie-Francine Moens 和 Tinne Tuytelaars。让运动栩栩如生：将静态图像转化为动态视频。载于《欧洲计算机视觉会议》，2024年。1[16] Mingxiao Li、Mang Ning 和 Marie-Francine Moens。一致的故事生成：释放之字形采样的潜力。载于《神经信息处理系统进展》，2025年。[17] Mingxiao Li、Tingyu Qu、Tinne Tuytelaars 和 Marie-Francine Moens。迈向更准确的个性化图像生成：解决过拟合和评估偏差。arXiv预印本 arXiv:2503.06632，2025年。1[18] Tianhong Li 和 Kaiming He。回归基础：让去噪生成模型执行去噪。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2026。2, 12[19] Yujia Li、Kevin Swersky 和 Richard Zemel。生成矩匹配网络。载于《第32届国际机器学习会议论文集》，2015年。2[20] Zhuang Liu、Hanzi Mao、Chao-Yuan Wu、Christoph Feichtenhofer、Trevor Darrell 和 Saining Xie。面向2020年代的卷积网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年。3[21] Yiyang Lu、Susie Lu、Qiao Sun、Hanhong Zhao、Zhicheng Jiang、Xianbang Wang、Tianhong Li、Zhengyang Geng 和 Kaiming He。利用像素均值流实现单步无潜变量图像生成。载于《第43届国际机器学习会议论文集》，第81265–81277页，2026。1, 2, 6[22] Mang Ning、Mingxiao Li、Jianlin Su、Albert Ali Salah 和 Itir Onal Ertugrul。阐明扩散模型中的曝光偏差。载于《国际表征学习会议》，2024年。1, 2[23] Mang Ning、Mingxiao Li、Jianlin Su、Haozhe Jia、Lan miao Liu、Martin Bene˘s、Wenshuo Chen、Albert Ali Salah 和 Itir Onal Ertugrul。DCTdiff：DCT空间中图像生成建模的独特特性。载于《第42届国际机器学习会议论文集》，2025年。2

[24] Mang Ning、Mingxiao Li、Le Zhang、Lanmiao Liu、Matthew B. Blaschko、Albert Ali Salah 和 Itir Onal Ertugrul。频谱匹配：在潜在扩散中实现卓越可扩散性的统一视角。arXiv预印本 arXiv:2603.14645，2026。3[25] Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby、Mahmoud Assran、Nicolas Ballas、Wojciech Galuba、Russell Howes、Po-Yao Huang、Shang-Wen Li、Ishan Misra、Michael Rabbat、Vasu Sharma、Gabriel Synnaeve、Hu Xu、Herv´e Jegou、Julien Mairal、Patrick Labatut、Armand Joulin 和 Piotr Bojanowski。DINOv2：无需监督地学习稳健的视觉特征。arXiv预印本 arXiv:2304.07193，2023年。4[26] Woobin Park、Jaeyeon Kim、Yilun Du、Sitan Chen 和 Bohyung Han。我们是否过度思考单步生成？arXiv预印本 arXiv:2610.20331，2026。1, 2, 3, 4, 6, 11, 12[27] William Peebles 和 Saining Xie。基于 Transformer 的可扩展扩散模型。载于《IEEE/CVF国际计算机视觉会议论文集》，2023年。1[28] Fei Peng、Junqiang Wu、Haoxian Tan、Jun Zhou、Jie Hu、Xiaoming Wei、Jie Guo 和 Xiu Li。高效混合蒸馏：协同利用得分和对抗目标函数实现一步扩散。arXiv预印本 arXiv:2611.03017，2026。3[29] Jaskirat Singh、Xingjian Leng、Zongze Wu、Liang Zheng、Richard Zhang、Eli Shechtman 和 Saining Xie。表征对齐中什么更重要：全局信息还是空间结构？arXiv预印本 arXiv:2512.10794，2025年。2, 3[30] Yang Song、Prafulla Dhariwal、Mark Chen 和 Ilya Sutskever。一致性模型。载于《第40届国际机器学习会议论文集》，2023年。1, 2[31] Hugues Van Assel、Edward De Brouwer、Saeed Saremi、Gabriele Scalia 和 Aviv Regev。在重建空间中生成，在语义空间中匹配：单步生成的传输几何。arXiv预印本 arXiv:2606 00514，2026 1, 2, 3 . . [32] Runqian Wang 和 Kaiming He。扩散与分散：基于表征正则化的图像生成。arXiv预印本 arXiv:2506.09027，2025年。3[33] Ziqiao Wang、Wangbo Zhao、Yuhao Zhou、Zekai Li、Zhiyuan Liang、Mingjia Shi、Xuanlei Zhao、Pengfei Zhou、Kaipeng Zhang、Zhangyang Wang、Kai Wang 和 Yang You。REPA 的作用并非始终如一：提前停止、整体式对齐大幅提升扩散训练。ar Xiv预印本 arXiv:2505.16792，2025年。2, 3, 7[34] Jiawei Yang、Zhengyang Geng、Xuan Ju、Yonglong Tian 和 Yue Wang。用于视觉生成的表征 Fr´echet 损失。arXiv预印本 arXiv:2604.28190，2026。2, 5, 12[35] Jingfeng Yao、Bin Yang 和 Xinggang Wang。重建与生成：驾驭潜在扩散模型的优化困境。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2025年。3[36] Tianwei Yin、Micha¨el Gharbi、Taesung Park、Richard Zhang、Eli Shechtman、Fredo Durand 和 William T. Freeman。改进的分布匹配蒸馏，实现快速图像合成。arXiv预印本 arXiv:2405.14867，2024年。1, 2

[37] Tianwei Yin、Micha¨el Gharbi、Richard Zhang、Eli Shecht- man、Fredo Durand、William T. Freeman 和 Taesung Park。采用分布匹配蒸馏的一步扩散。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，2024年。 1, 2[38] Sihyun Yu、Sangkyung Kwak、Huiwon Jang、Jongheon Jeong、Jonathan Huang、Jinwoo Shin 和 Saining Xie。用于生成的表示对齐：训练扩散 Transformer 比你想象的更容易。在 国际学习表征会议，2025年。 2, 3[39] Richard Zhang、Phillip Isola、Alexei A. Efros、Eli Shecht- man 和 Oliver Wang。深度特征作为感知度量的非凡有效性。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，2018。 3[40] Boyang Zheng、Nanye Ma、Shengbang Tong 和 Saining Xie。采用表示自编码器的扩散 Transformer。 *arXiv*预印本 *arXiv:2510.11690*，2025年。 3[41] Linqi Zhou、Stefano Ermon 和 Jiaming Song。归纳矩匹配。在第*42*届国际机器学习会议论文集，2025年。 2

对齐配对项，而非样本：面向更快、更优单步生成器的表征对齐补充材料

本补充材料包含命题 1 的证明（第 A节）、实现与评估细节（第 B 和 C节）、计算量统计（第 D节）、完整结果表（第 E节）、附加分析（第 F节）以及未筛选样本（第G节）。所有数值均直接读取自每次运行所存储的评估输出；可根据这些输出重新生成所有表格和图表的脚本属于代码发布包。

A. 命题 1 的证明

令 X 为类别为 c 的生成器输入，$T \in \mathbb{S}^{d-1}$ 为目标标记，且 $u : \mathcal{X} \to \mathbb{S}^{d-1}$ 为任意可测预测。记 $\mu(X) = \mathbb{E}[T \mid X]$ 和 $m_c = \mathbb{E}[T \mid c]$。

(i) 根据塔式性质，$\mathcal{R}(u) = 1 - \mathbb{E}\big[u(X)^\top \mathbb{E}[T \mid X]\big] = 1 - \mathbb{E}[u(X)^\top \mu(X)]$. 由于 $\|u(X)\| = 1$，由柯西–施瓦茨不等式可得 $u(X)^\top \mu(X) \le \|\mu(X)\|$ 逐点成立，等号成立当且仅当 $u(X) = \mu(X)/\|\mu(X)\|$ 只要 $\mu(X) \neq 0$（当 $\mu(X) = 0$ 时，任意单位向量都是最优的）。取期望可得 $\mathcal{R}(u) \ge 1 - \mathbb{E}\|\mu(X)\|$，并由 $u^\star$ 取到。

*(ii)*若 $T \perp X \mid c$，则 $\mu(X) = \mathbb{E}[T \mid X] = \mathbb{E}[T \mid c] = m_c$，其仅通过 c取决于 X。根据 (i)，对于类别 c 的所有 X（若$m_c \neq 0$），最优预测为 $u^\star(X) = m_c/\|m_c\|$，风险为 $1 - \mathbb{E}_c\|m_c\|$。

(iii) 由于 c 是 X 的函数（类别标签是输入），$m_c = \mathbb{E}[\mu(X) \mid c]$。范数是凸函数，因此根据条件詹森不等式 $\mathbb{E}[\|\mu(X)\| \mid c] \ge \|\mathbb{E}[\mu(X) \mid c]\| = \|m_c\|$。当且仅当在每个类别内，向量 $\mu(X)$几乎必然是某一共同方向的非负倍数时，等号成立，即，当且仅当 $u^\star$ 在每个类别内为常数。对 c 取平均可得$\mathcal{A} \ge 0$，并满足所述等号条件。□

备注。(1) 该命题仅涉及对齐项本身在总体上的最优值。在训练中，对齐损失会与重建损失或 OT损失结合，并通过投影器 P_ψ发挥作用。尽管如此，对于非配对目标，其相对于生成器特征的梯度仍指向一种经过投影后类别常数的特征配置；这就是主论文讨论的吸引子。(2) 对于跨标记取最大值的余弦相似度（用于我们的 10 种生成路径配置），$\mathcal{R}$ 会替换为 $\mathbb{E}[1 - \max_q u_p^\top T_q]$。对于随机目标或池化目标，给定 c 后，目标集合仍独立于 X，因此最小化解仍然是类别常数。我们使用 1024 张真实 ImageNet 图像的 DINOv2 令牌，测量常数预测的接近程度。单个类别均值方向的最大余弦相似度达到 0.60；由于它是在训练样本上拟合的，因此这是一个偏乐观的结果。使用逐图像标记选择的训练线性探针达到 0.63，而单个全局方向达到 0.37。因此，类别常数预测器能够达到大部分可实现的最大余弦相似度。(3) 对于配对路径，$\mathcal{A} > 0$ 是一个经验性结论。在发布的 Priming-B（160 个训练轮次，无对齐）中，使用 x_s (s=0.8) 的第8块特征训练岭回归探针，可以恢复干净图像的 DINOv2 令牌，令牌级余弦相似度为 0.36。将图像配对打乱后，余弦相似度降至 0.18。相较于这一打乱后的基线，生成器特征携带了 $+0.185$ 可恢复的目标结构；而在 x_s 的原始像素上训练的探针则为 $+0.041$ （80/20 图像级划分，自助法标准差为 0.004）。

B. 实现细节

生成器（来自 *[26]*）	
B / L	16 / 32 个块，宽度为 768 / 1024；115.1M / 405.1M 个参数
标记	图像块 16（16 × 16 网格）+ 8 个类别令牌，二维旋转位置编码
预热（第一阶段）	
扰动	$x_s = 0.2\,x_0 + 0.8\,\epsilon$（固定 $s = 0.8$）
Loss	自适应像素均方误差 + 0.4 LPIPS + 0.1 ConvNeXt
优化器	Muon（动量 0.95），学习率 10^{-3} 恒定，无权重衰减
批次 / 指数移动平均	1024 / 0.9999；脑浮点 16
训练轮次	160（200、160 个更新步数）；预算 10–200
数据增强	中心裁剪 256^2, 随机水平翻转
***RePrime* 对齐**	
教师	DINOv2-L/14 (lvd142m)，冻结，FP32
教师输入	x_0 调整至 $224^2 \to 16 \times 16$ 图块令牌
生成器特征	区块 5（B，默认）、8（B，最佳）、10（L）
投影器	线性层(D,1024)–GELU–线性层(1024,1024)
投影器优化器	AdamW，学习率 10^{-4}
Loss	逐令牌 $1-$余弦，$\lambda = 0.5$，整体 Priming
后训练（阶段 *2*，从 *[26]* 开始）	
更新步数	25k（扩展运行时为 50k）
优化器	AdamW，学习率 10^{-5}（预热 125 步），裁剪阈值 1.0
EMA	0.999
批次	16 个类别 × （64 + 64 停止梯度）样本，每个类别 128 个真实样本
OT	0.75 最优传输$_{0.001}$ + 0.25最优传输$_{0.01}$，对数 Sinkhorn
特征	ConvNeXt + ResNet-MAE 堆叠，系数 5
RePrime-J	
最优传输频率	每 $k = 2$外部更新
最优传输权重	在线性递增 $0 \to 1$前 25k 个更新步数内
调度	200,160 次联合更新（扩展运行时为 240,192 次）；可选精炼：阶段 2，更新步数为 25k

表 S1。**超参数**。RePrime 中未列出的所有内容在 Priming 和 RePrime 中均相同。

生成路径对齐。 所有生成路径运行均从发布的 Priming-B 160 训练轮次初始化开始，并使用标准后训练方案进行 2.5k 次更新。对齐损失通过与上述形式相同的投影器作用于区块 8。除非另有说明，它在 250 次更新后启用，并在线性地经过 100 次更新逐步增加至 $\lambda = 1$。目标为：(i) 来自最优传输参考集、与随机真实图像同类别的真实图像的 DINOv2 令牌；(ii) 一组来自所有同类参考图像的令牌，并采用标记最大匹配；(iii) 对每个样本而言，具有最大 OT 计划质量的参考图像的 DINOv2 令牌；(iv) 样本自身的 DINOv2 令牌（停止梯度）；以及 (v)

同一噪声下 EMA 教师生成样本的令牌。对于 ConvNeXt 变体，目标是冻结的 OT ConvNeXt 的阶段 2 令牌。最大余弦相似度将每个预测令牌与其最相似的目标令牌进行匹配，而逐令牌匹配则将令牌按相同位置进行匹配。这次 2.5k 次更新实验中的所有运行，包括无对齐基线和配对路径对应项，都使用了 ImageNet 训练集的较早导出版本，其打包方式不同于完整预算实验中使用的官方版本。因此，我们仅在彼此之间进行比较。

联合训练。 OT 步骤在 Priming 循环中使用后训练机制：每隔一次更新，针对 16 个类别的 $\times$ 64 样本（以及一个停止梯度的负批次）执行一次前向传播/反向传播。OT 梯度单独进行全规约，并在执行一次优化器步之前加到重建梯度上。联合训练分阶段进行（40 个训练轮次，然后继续训练至 160 和 192 个训练轮次），各阶段使用相同的超参数。每个阶段初始升温后，记录的 OT 权重均为 1.0。

C. 评估协议

FID 和FD_r^5。 我们使用 [26]的官方评估脚本，采用 50k 个样本、一次网络评估、不使用引导，并使用 EMA 权重。FID 使用 JiT ImageNet-256 参考统计量 [18]。FD_r^5 对 Inception、DINOv2、ViT-MAE、SigLIP 和 CLIP（CLS）特征中的 Fr´echet 距离取平均值，并分别使用验证集参考值进行归一化 [26, 34]。文中报告的数值在注明时取三个采样随机种子的平均值。对于训练 160 个周期的 RePrime-B，同一检查点评估两次，每次使用三个随机种子，并对全部六次评估结果取平均值。

FD_r^6，精确率、召回率。 训练时评估器在 50k 个样本上计算六种 FD（Inception/ADM、CLIP、ConvNeXt、DINOv2、MAE、SigLIP），并报告

$$\mathrm{FD}_r^6 = \tfrac{1}{6}\Big(\tfrac{\mathrm{FD}_{\mathrm{Inc}}}{1.68}+\tfrac{\mathrm{FD}_{\mathrm{CLIP}}}{5.6}+\tfrac{\mathrm{FD}_{\mathrm{CNX}}}{56.87}+\tfrac{\mathrm{FD}_{\mathrm{DINO}}}{14.19}+\tfrac{\mathrm{FD}_{\mathrm{MAE}}}{0.04}+\tfrac{\mathrm{FD}_{\mathrm{SigLIP}}}{0.6}\Big) \tag{S1}$$

以及精确率和召回率（$k = 3$、Inception 特征、ADM ImageNet-256 参考批次）。该评估器每进行 1.25k 次后训练更新，就会分别对在线权重和指数移动平均权重运行一次。其 Inception FD 并非官方 FID。

D. 计算量核算

我们遵循 [26] 的 EFLOP 核算方式。根据报告的数据，Priming-B 在 80/160 个训练轮次下分别消耗 85.75/124.18 EFLOPs，据此可得，每个 Priming 训练轮次消耗 0.480 EFLOPs，后训练（25k 个更新步数）消耗 47.3 EFLOPs。对于 L，这两个数值分别为 1.096 和 63.6 EFLOPs。DINOv2-L/14 的一次前向传播（24 层，宽度 1024，257 个标记）每张图像消耗 $2 \cdot 24\,(12Td^2 + 2T^2d)$ $= 162$十亿次浮点运算，即，每个训练轮次消耗 0.207 EFLOPs。这使 Priming 每个训练轮次的计算量分别增加 43%（B）和 19%（L）。一次 OT 更新消耗 $47.3/25,\ 000 = 1.89$ 千万亿次浮点运算，即，约为一次 B预热更新的 $4.9\times$ 倍，因此，包含 100k 次 OT 更新的联合运行会额外消耗 189 EFLOPs。我们的实现采用 FP32 评估教师，且未进行编译。在 8×H200 上测得，RePrime 每次 B预热更新的实际运行时间为 1.23 秒，而编译基线为 0.66– 0.72 秒。教师特征仅取决于图像及其翻转版本，因此可以预先计算一次。

E. 完整结果

本节列出主论文图表背后的数据，并概述各表所展示的内容。

预热预算（表 S2）。在所有七种预算下，RePrime 在每项指标上均优于 Priming：FID、FD_r^5、FD_r^6和召回率。相对FD_r^5 增幅在 10 个训练轮次时最大（-16%），到 200 个训练轮次时仍为 -11%。RePrime 训练 80 个训练轮次的表现与 Priming 训练 160 个训练轮次相当（FD_r^5 11.24 对比 11.27）；RePrime 训练 120 个训练轮次则在四项指标上都胜过 Priming 训练 200 个训练轮次。每种预算下召回率都有所提升，因此这一增益并非来自样本集中化。教师的计算开销使 RePrime 的各个点在图中向右移动，但 RePrime 训练 80 个训练轮次仍以少 2.5% 的总计算量胜过 Priming 训练 120 个训练轮次。

生成路径配置（表 S3）。每一行的召回率都远低于基线（0.149–0.252 对比 EMA 的 0.452）。17 行中有 16 行的精确率达到或超过基线，IS 也常常升高。样本变得更典型，而非单个样本变差，这正是命题 1 所预测的原型吸引现象。这一模式不受目标构造、目标空间、损失形式、损失权重或对齐窗口影响。与样本本身或 EMA 教师对齐，效果与和随机图像对齐相似。对表现最差的运行继续进行 2.5k 次纯 OT 更新，也只能恢复少部分损失的召回率。**各空间距离**（表 S4）。我们的主要结论不应取决于教师自身的特征空间。去掉 DINOv2 列后，平均值仍分别下降 8.6%（B-160）、10.8%（B-200）和 12.7%（L）。在 L 下，RePrime 改善了所有空间。在 B 下，除了训练 160 个训练轮次时 CLIP 的表现持平（22.14 对比 22.15）之外，RePrime 改善了所有空间；训练 200 个训练轮次时，CLIP 也有所改善。相对增益最大的是 SigLIP（16–19%），而训练阶段和教师都未使用这一空间。经过联合训练和精炼后，RePrime-J 改善了除 CLIP 之外的所有空间。

从 40 个训练轮次开始的长时间后训练（表 S5）。在所有后训练时长下，RePrime 初始化都领先；其在 FD_r^6 上的优势虽有所缩小，但仍然存在（1.25k 更新步数时为 3.50，25k 时为 2.02）。RePrime 经过 2.5k 更新步数后达到 13.71，与 Priming 经过 22.5k 更新步数后的结果相当，$9\times$ 节省了成本最高阶段的开销。每个步骤的召回率也都更高。因此，后训练

承接 Priming 阶段对齐所取得的成果，而不是从较弱的起点奋起直追。

联合训练（表格 S6 和 S7）。原生 RePrime-J 在FD_r^5 上每个检查点都优于原生联合训练，其原生 FID 间隔从 50k 更新步时的 1.16 稳步缩小至 240k 更新步时的 0.17。对齐深度会在两个指标之间进行权衡：FD_r类指标偏好区块 5，而原生 FID 则随深度增加而改善。由于生成路径从一开始就参与训练，更深的接口可能会使噪声路径在像素解码前可用于适应的块更少。两阶段 RePrime 则在FD_r^6 和召回率方面都偏好区块 8（主论文，表 2）。

F. 附加分析

F.1. 机制：接口与语义路由

定义。我们使用 1,024 张 ImageNet 图像（16 个类别 $\times$64）。在 数据模式下，生成器接收 $x_s = 0.2\,x_0 + 0.8\,\epsilon$；在 噪声模式下，它接收 $z \sim \mathcal{N}(0,\ I)$ 以及相同的标签。对于每个块，我们对令牌特征进行均值池化。噪声模式与数据模式的间隔是 Fr´echet 距离除以数据模式协方差的迹；接口是 2.5k 次最优传输更新后间隔最小的块，但排除最后一个块，因为该块负责生成像素输出，并且在每个模型中都有较小的间隔。扩散比是噪声模式特征的类内协方差的迹除以数据模式特征的类内协方差的迹。对于语义路由，我们使用数据模式特征拟合一个岭回归探针，以预测干净图像的 DINOv2 嵌入，将其应用于噪声模式特征，并报告 R^2，其参照是由该噪声生成的图像的 DINOv2 嵌入。CLIP 探针则使用数据模式特征拟合干净图像的留出集 CLIP 嵌入（第 F.2节）。FD_r^6 主论文图 5b 中的数值来自 2.5k 次最优传输更新后的在线权重。

后训练只是小幅扰动。在 2.5 万次最优传输更新期间，每个块的变化幅度仅为其范数的 2–4%，而且两种配方几乎完全相同（图 S1d ）。因此，后训练所利用的结构必定来自 Priming。

接口在后训练过程中仍然存在。对于 160 训练轮次这一组（图 S1a），后训练前，RePrime 在区块 5 的间隔为 4.47，而 Priming 为 28.35。经过 2,500 次更新后，RePrime 的最小值位于区块 5（0.86），而 Priming 的最小值位于区块 10（1.25）。经过 25,000 次更新后，两者的最小值分别漂移至 1.14 和 2.23。在 Priming 的预算范围内，从 10 个训练轮次起，RePrime 的接口便位于区块 5，间隔为 4.0–4.5；相比之下，基线的最小值仅从 9.96 降至 6.21，且始终位于区块 10–12。

对应关系对照。 表 S8 和图S1(b,c) 报告了主论文对应关系对照中四个 10 个 epoch 的实验组的诊断结果。三个对齐分支都在区块 5 形成接口，而随机目标实验组在后训练前的间隔最小（2.74，对比置换后的和 4.06 的自身图像）。然而，其特征不包含实例语义：数据模式 CLIP 探针的结果与无对齐相同；经过 2.5k 次最优传输更新后，其噪声输入几乎没有通过接口进行路由，扩散比也很低，也就是说，同一类别的噪声编码彼此靠得很近。置换后的标记保留了每图像令牌集合，并恢复了大部分路由和扩散，与其处于中间水平的 FD_r^6 相符。经过 5k 次最优传输更新后，这一模式依然存在：在区块 5，无对齐、随机目标、置换后的标记和自身图像的间隔分别为 6.43、1.84、1.10 和 1.14；路由分别为 0.03 、 -0.03 、 0.08 和 0.19；扩散比分别为 0.19、0.26、0.88 和 0.76。

生成路径对齐。 在生成路径（区块 8）上进行对齐后，对齐块处的间隔较小（0.47–0.61），因为对齐使噪声特征 2–4$\times$ 的类别均值进一步远离数据均值，同时缩小了类内扩散（随机目标为 0.13，未对齐时为 0.41）。区块 5 处的路由降至 0.10–0.16（未对齐时为 0.20），而 DINOv2 空间中的样本多样性从 0.376 降至 0.311–0.345。因此，仅凭间隔较小并不能作为诊断依据：在主论文图 5b 的 16 个模型中，间隔与 FD_r^6 的斯皮尔曼相关为 0.00，而路由达到 -0.90（ 27 个组内配对中有 21 个一致）。

未决问题。 生成路径表现最弱的方案（OT 匹配目标，λ=0.05）召回率下降（0.252），但编码层面并未出现崩溃（多样性为 0.372，扩散为 0.44，其接口处的路由为 0.34）。因此，其覆盖损失必定源于后续块，或源于将编码解码为像素的过程；我们的块池化诊断无法确定具体原因。

F.2. 表示探针

我们将相同的 1,024 张带噪真实图像（16 个类别 $\times$ 64，$s = 0.8$，固定噪声）输入两种配方在训练 10、40 和 160 个轮次后的 Priming 检查点，并使用每类 48 张训练图像和 16 张验证图像，将池化后的块ℓ 特征通过岭回归探针拟合到干净图像池化后的 DINOv2 和 CLIP 嵌入（图 S2）。RePrime 特征在所有计算预算下、在两种空间中（包括训练中完全未使用的 CLIP 空间）都具有更强的预测能力：在区块 5，RePrime 的 CLIP R^2 在训练 10/40/160 个轮次后分别为 0.34/0.44/0.46，而 Priming 分别为 0.32/0.37/0.36。优势在对齐块处达到峰值。训练 40 个轮次后，RePrime（区块 5 上为 0.435）已超过训练 160 个轮次的基线在其最佳区块（区块 10）上的表现（0.414）。这与后续最优传输阶段更快收敛的情况一致，该阶段正是基于这些特征运行的。对于相同输入，训练 160 个轮次后的重建在 DINOv2 空间中与其自身的干净图像比与其类别原型更接近（余弦相似度分别为 0.81 和 0.73），也就是说，配对路径输入携带了关于对齐目标的实例信息（$\mathcal{A} > 0$见命题 1）。

	Priming-B					RePrime-B				
Ep.	EFLOPs	FID	FD_r^5	FD_r^6	Rec.	EFLOPs	FID	FD_r^5	FD_r^6	Rec.
10	52.1	2.551	21.63	18.27	0.556	54.2	**2.157**	**18.23**	**15.39**	**0.576**
20	56.9	1.993	17.79	14.99	0.581	61.1	**1.709**	**15.38**	**12.96**	**0.596**
40	66.5	1.703	15.22	12.80	0.604	74.8	**1.507**	**12.96**	**10.93**	**0.614**
80	85.8	1.495	12.90	10.86	0.614	102.3	**1.404**	**11.24**	**9.43**	**0.623**
120	105.0	1.415	11.82	9.97	0.619	129.8	**1.349**	**10.76**	**9.02**	**0.627**
160	124.2	1.373	11.27	9.51	0.620	157.3	**1.303**	**10.20**	**8.61**	**0.639**
200	143.4	1.353	10.91	9.21	0.625	184.8	**1.283**	**9.66**	**8.19**	**0.631**

表 S2. **Priming 预算扫描 (B)**。每个模型均经过 25k 次更新的后训练。FID/FD_r^5 为单随机种子结果，但 160 和 200 个训练轮次除外，这两个轮次采用与主论文表 1 相同的随机种子平均值（三个随机种子；RePrime 在 160 个训练轮次时评估六次）。FD_r^6 和召回率使用 EMA 权重。80 个训练轮次的 Priming 模型为发布的检查点；其他 Priming 预算来自我们对官方方案的重新训练（10–160 个训练轮次：一次运行并设置里程碑；200 个训练轮次：单独运行）。

					EMA				在线	
目标	目标空间	Loss	λ	窗口	FD_r^6	Rec.	精确率	IS	FD_r^6	Rec.
无对齐（基线）					14.71	0.452	0.787	270	13.80	0.494
最优传输匹配的真实样本	DINOv2	最大余弦相似度	1	250–2.5k	16.86	0.200	0.824	290	17.22	0.246
同类随机样本	DINOv2	最大余弦相似度	1	250–2.5k	16.71	0.203	0.826	291	17.18	0.255
同类令牌池	DINOv2	最大余弦相似度	1	250–2.5k	17.14	0.189	0.839	297	18.30	0.249
自身样本（停止梯度）	DINOv2	最大余弦相似度	1	250–2.5k	16.90	0.201	0.820	286	17.78	0.257
EMA 教师样本	DINOv2	最大余弦相似度	1	250–2.5k	17.93	0.182	0.830	290	19.42	0.227
最优传输匹配的真实样本	ConvNeXt（OT）	最大余弦相似度	1	250–2.5k	16.28	0.195	0.835	297	16.02	0.256
同类随机样本	ConvNeXt（OT）	最大余弦相似度	1	250–2.5k	16.43	0.193	0.836	299	16.05	0.268
同类令牌池	ConvNeXt（OT）	最大余弦相似度	1	250–2.5k	16.87	0.181	0.840	301	16.55	0.256
自身样本（停止梯度）	ConvNeXt（OT）	最大余弦相似度	1	250–2.5k	16.46	0.192	0.830	299	16.11	0.261
EMA 教师样本	ConvNeXt（OT）	最大余弦相似度	1	250–2.5k	16.56	0.186	0.833	299	16.11	0.257
最优传输匹配的真实样本	DINOv2	逐令牌	1	250–2.5k	19.20	0.171	0.843	268	22.63	0.198
同类随机样本	DINOv2	逐令牌	1	250–2.5k	19.62	0.149	0.858	278	24.08	0.200
最优传输匹配的真实样本	DINOv2	逐令牌	0.25	250–2.5k	16.07	0.230	0.804	261	16.98	0.244
最优传输匹配的真实样本	DINOv2	逐令牌	0.1	250–2.5k	15.39	0.252	0.790	266	14.94	0.261
最优传输匹配的真实样本	DINOv2	逐令牌	0.05	250–2.5k	14.97	0.252	0.793	270	14.28	0.276
+ 2.5k 纯 OT 更新（总计 5k）					13.27	0.290	0.781	263	13.51	0.299
最优传输匹配的真实样本	DINOv2	逐令牌	1	250–750	14.92	0.242	0.799	268	14.07	0.271
最优传输匹配的真实样本	DINOv2	逐令牌	0.25	250–750	14.99	0.252	0.782	268	13.98	0.287
配对路径：RePrime 初始化，后训练阶段无对齐					**12.89**	**0.469**	0.788	269	**12.36**	**0.502**

表S3。**所有生成路径对齐配置**（B-160初始化，2500次训练后更新）。所有配置都会使召回率降低 0.20–0.30，对指数移动平均权重和在线权重均如此。缩进行是在 λ=0.05 运行基础上继续进行 2500 次最优传输更新且不进行对齐；召回率仅恢复至 0.290，远低于 2500 次更新的基线（0.452）。最后一行是在相同方案下的配对路径对应项：采用 RePrime-B 初始化，并在不进行任何对齐的情况下进行后训练。

模型	Inception	CLIP	SigLIP	ViT-MAE	DINOv2†	FD_r^5 无 DINOv2
Priming-B (160)	0.836	22.14	20.07	6.79	6.49	11.27	12.46
RePrime-B (160)	0.780	22.15	16.79	5.82	5.47	10.20	11.38
Priming-B (200)	0.809	21.61	19.27	6.61	6.27	10.91	12.07
RePrime-B（200）	0.772	20.75	16.03	5.51	5.25	9.66	10.76
Priming-L（160）	0.738	15.13	12.83	4.05	4.32	7.41	8.19
RePrime-L（160）	0.730	13.93	10.44	3.48	3.57	6.43	7.14
Joint-B + 精修	0.768	22.22	21.51	6.69	6.42	11.52	12.80
RePrime-J-B +精修	0.748	23.18	17.80	5.78	5.52	10.60	11.88

表 S4。**官方 FD 评估中按空间归一化的 Fréchet 距离**（越低越好），取所有采样随机种子的平均值（每个模型使用三个，训练 160 个轮次的 RePrime-B 使用六个）。†DINOv2 是我们的教师，因此最后一列给出不包含它时重新计算的平均值。

	Priming 40 轮		RePrime 40 轮	
最优传输更新	FD_r^6	Rec.	FD_r^6	Rec.
1.25k	18.11	0.504	14.61	0.532
2.5k	16.64	0.538	13.71	0.566
5k	15.89	0.563	13.67	0.591
10k	14.60	0.586	12.76	0.607
15k	14.10	0.583	12.21	0.606
20k	13.86	0.597	11.92	0.619
22.5k	13.71	0.595	12.00	0.622
25k	13.60	0.600	11.58	0.608

表 S5. **40 轮初始化，长时间后训练**（在线权重）。经过 2.5k 次更新步数后，RePrime 达到 Priming 经过 22.5k 次更新步数后的效果。

F.3. 最优传输计划的选择性有多高?

使用 OT 匹配目标进行生成路径对齐，假设该计划将每个样本与语义上相符的对象配对——对应的真实图像。我们在后训练开始时对此进行测试。对于上述每个 Priming 检查点，我们从纯噪声中为每个类别生成 64 个样本，并计算第二阶段 Sinkhorn 计划

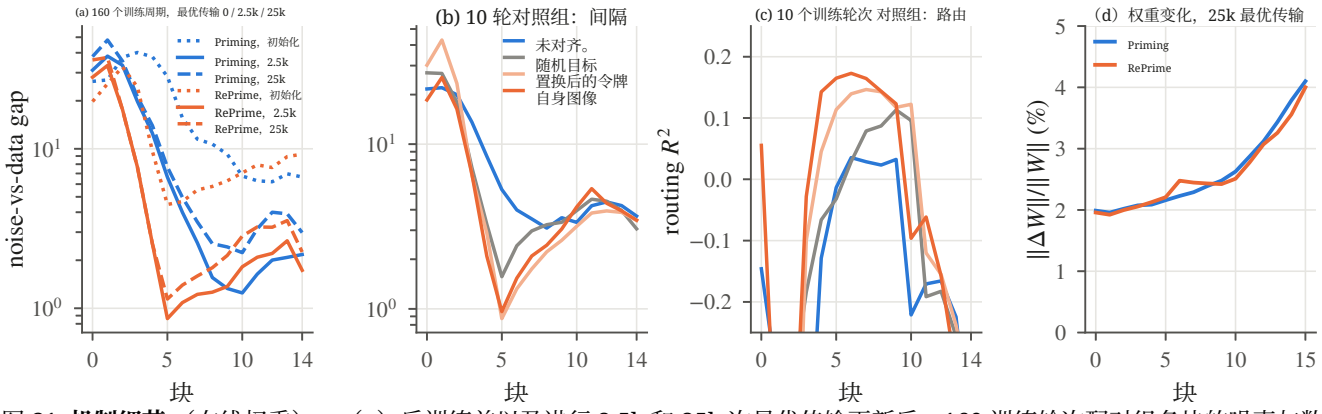


图 S1. **机制细节**（在线权重）。（a）后训练前以及进行 2.5k 和 25k 次最优传输更新后，160 训练轮次配对组各块的噪声与数据差距。（b,c）进行 2.5k 次最优传输更新后，10 训练轮次对应关系对照组各块的差距和语义路由；所有对齐分支都在区块 5 形成接口，但只有实例配对目标会通过该接口传递语义。（d）进行 25k 次最优传输更新期间各块的相对权重变化。

更新步数	联合（FID / FD_r^5）	RePrime-J（FID / FD_r^5）
50k	3.68 / 16.71	4.84 / 13.96
60k	3.53 / 15.53	4.58 / 13.28
70k	3.27 / 14.70	4.30 / 12.67
80k	3.06 / 14.12	3.99 / 12.12
90k	–	3.84 / 11.74
100k	2.73 / 13.04	3.72 / 11.45
110k	2.63 / 12.63	3.62 / 11.20
120k	2.59 / 12.44	3.52 / 10.98
130k	2.54 / 12.10	3.39 / 10.68
140k	2.35 / 11.95	3.11 / 10.46
150k	2.25 / 11.69	3.03 / 10.18
160k	2.24 / 11.57	2.85 / 10.08
170k	2.16 / 11.41	2.74 / 9.97
180k	2.10 / 11.30	–
190k	2.08 / 11.17	–
200k	2.06 / 11.00	2.47 / 9.75
240k	2.05 / 10.54	2.22 / 9.33

表 S6。**原生联合模型**直接从噪声进行评估（指数移动平均权重，单个随机种子）。两次运行遵循相同的调度；超过 200k 的检查点来自继续训练至 240k 更新步数。

块	FD_r^6 (+1.25k 最优传输，在线)	原生 FID	原生 $FD^5{}_r$
3	13.06	6.12	15.75
5	12.44	4.84	13.96
8	12.87	4.50	14.63
12	13.41	3.90	15.85

表 S7。**RePrime-J 中的对齐深度**（40 个训练轮次）。FD_r 型指标偏好区块 5，而原生 FID 随深度增加持续改善。两阶段 RePrime 偏好区块 8（主论文，表 2）。

在 ConvNeXt/MAE 匹配空间中，将其与每个类别的 128 个真实参考样本进行匹配（对其特征级问题取平均），并测量每个样本与以下对象之间的 DINOv2 余弦相似度：(i) 其计划加权参考样本，(ii) 计划质量最大的参考样本，(iii) 同一类别中均匀随机选取的参考样本，以及 (iv) 其最近参考样本。在全部六个检查点中，(i) 比 (iii) 至多高出 0.002，(ii) 比 (iii) 至多高出 0.012（例如，在 160 个训练轮次时为 0.221 对 0.213），

目标（10 个训练轮次）	间隔$_5$ 初始化 CLIP R_5^2		间隔$_5$	路由$_5$	扩散$_5$
无对齐	18.10	.360	5.30	−.01	.20
同类随机样本	2.74	.361	1.57	−.03	.26
自身的置换后标记	3.74	.406	0.87	.11	.85
自身图像	4.06	.413	0.96	.17	.76

表 S8。**区块 5 处对应关系对照的接口诊断**：后训练前（“初始化”，数据模式 CLIP 探针）以及后训练 2.5k 次最优传输更新后（间隔、路由、扩散）。在线 FD_r^6 在更新 2.5k 步后：24.24、23.89、21.75、20.28；指数移动平均 FD_r^6 在更新 5k 步后：22.10、21.62、19.94、18.10。

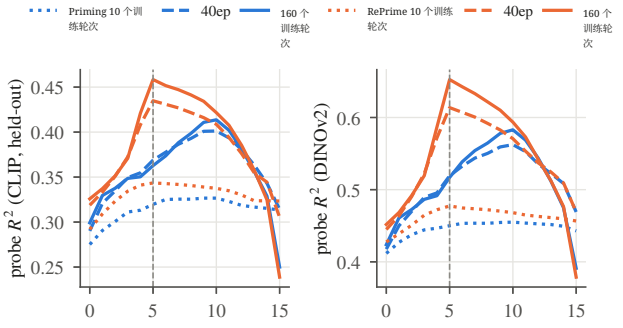


图 S2. **固定带噪真实输入（）的生成器块级特征的语义线性可读性**，使用岭回归探针预测干净图像的池化 CLIP（左侧，训练中从未使用）和 DINOv2（右侧）嵌入。在留出图像上进行验证 R^2；展示训练 10、40 和 160 个训练轮次后的预热检查点。

而 (iv) 达到 0.31–0.40。从语义编码器的角度看，该方案因此几乎是在类别内随机选择目标，这解释了为什么 OT 匹配目标与随机同类别目标表现相似（主论文，图 4；表 S3）。

F.4. 独立重跑与采样方差

采样方差很小：在同一检查点的三个采样随机种子下，标准差为 0.001– 0.022 FID 和 0.008–0.035 FD_r^5（主论文，表 1）。重新

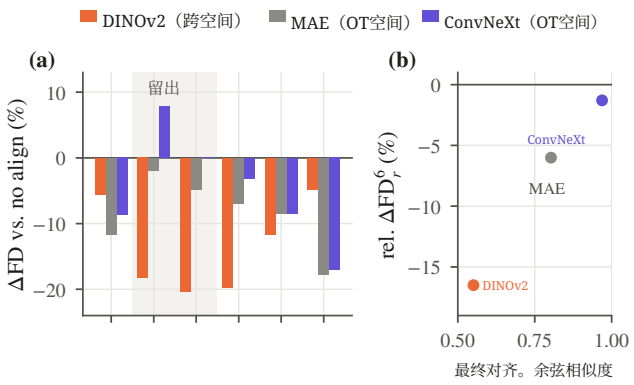


图 S3。**教师选择**（40轮预热、2500次训练后更新、指数移动平均）。（a）相对于无对齐，各特征空间中的 Fr´echet 距离变化。DINOv2 对齐使留出的 CLIP 和 SigLIP 空间提升了 18–20%，而对齐到 OT ConvNeXt 空间主要改善该空间，并且使 CLIP 变差。（b）易于对齐的教师带来的收益最小。

运行训练 会使结果变化更大，但其影响仍远小于对齐的影响。在 160 个训练轮次时，发布的 Priming- B 初始化和我们重新训练得到的初始化分别取得 FD^6_r 9.51/9.71 和 FID 1.373/1.367。两次独立的 RePrime 运行分别取得 8.61/8.47 和 1.303/1.302。配方间差异（0.9–1.2 FD^6_r）是同一配方内波动的四到六倍。

F.5. 教师空间

图 S3 按特征空间细分了主论文中 40 轮教师比较的结果。ConvNeXt 对齐使 ConvNeXt 距离提升了 17%，但使 CLIP 变差了 8%；相比之下，DINOv2 对齐使留出的 CLIP 和 SigLIP 空间提升了 18–20%；易于对齐的教师带来的收益最小（图 S3b）。

在完整预算下（160 个训练轮次 + 25k 次更新），对齐到最优传输特征空间并无帮助。ResNet-MAE 对齐得到 FD ^6_r 9.65，ConvNeXt 对齐得到 9.95（从相同初始化开始进行的两次后训练运行：9.95/9.95），而无对齐时为 9.51，使用 DINOv2 时为 8.61。即使在完整预算下，ConvNeXt 对齐在其自身 ConvNeXt 空间中的每空间距离也比基线更差（33.6 对 31.6），而 DINOv2 对齐则使其改善至 28.0。

F.6. 对齐余弦值与损失占比

经过 40 个预热轮次后，对齐余弦值分别为 0.55（DINOv2）、0.80（MAE）和 0.97（ConvNeXt）。相应的加权对齐项分别占预热损失总量的 13%、6% 和 1%。在生成路径上，ConvNeXt 空间中的目标几乎得到完美匹配（最大余弦相似度为 0.998），但召回率却骤降（表 S3）。因此，对齐得分几乎无法说明对齐的价值。

F.7. 联合训练曲线

图 S4 展示了联合模型在训练过程中的原生（无后训练）FID 和 FD^5_r，以及 $2\times2$40 个训练轮次的析因设计训练轮次（顺序式与联合式、有对齐与无对齐，每种设置之后均进行 1.25k 次更新步数的最优传输探测）。原生 RePrime-J 在约 110k 次联合更新步数后超过两阶段 Priming-B 的 FD^5_r，并在 200k 次更新步数后达到 9.75；其原生 FID 更高（无对齐时为 2.47→2.22，而非 2.06→2.05）。

G. 附加样本

图 S5将主论文中的图 4 扩展到四个类别和十二个噪声向量，并加入最弱的生成路径设置（λ=0.05）。图 S6 展示了在不同预热预算下固定噪声的结果。图S7 至S9 展示了 20 个类别的未筛选样本，每个类别使用六个固定噪声向量；没有选择或筛选任何样本。

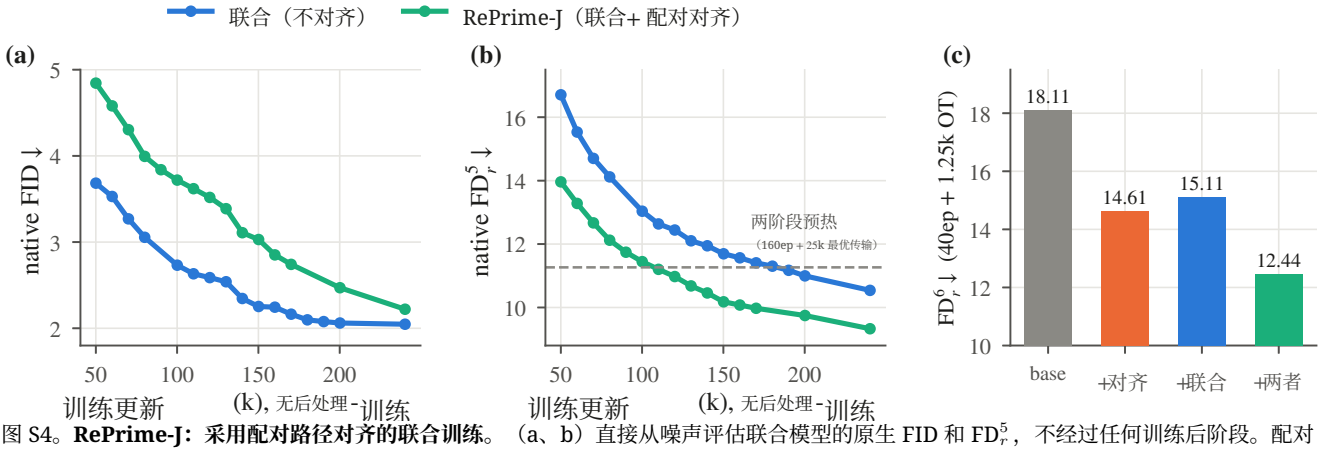


图 S4。**RePrime-J：采用配对路径对齐的联合训练**。（a、b）直接从噪声评估联合模型的原生 FID 和 FD_r^5，不经过任何训练后阶段。配对路径对齐以原生 FID 为代价，换来了显著更优的 FD_r^5；经过约 110k 次联合更新后，原生 RePrime-J 超越了两阶段 Priming-B 的 FD_r^5。（c）40 个训练轮次下的 2×2 析因设计（+1.25k 次 OT 更新，在线 FD_r^6）：对齐和联合训练分别带来提升，二者结合时效果最佳。



图 S5. **类别间的原型吸引** (金毛寻回犬、金刚鹦鹉、大熊猫、火山；每个块的每一行均使用相同的 12 个噪声向量；从发布的 Priming-B 初始化开始进行 2500 次训练后更新，但每个块的最后一行除外，该行从 RePrime 初始化开始)。在每个类别中，OT 匹配行与随机目标行几乎完全相同，即，OT 匹配图像充当类别级目标。这些模型的召回率和精确率列于表 S3。

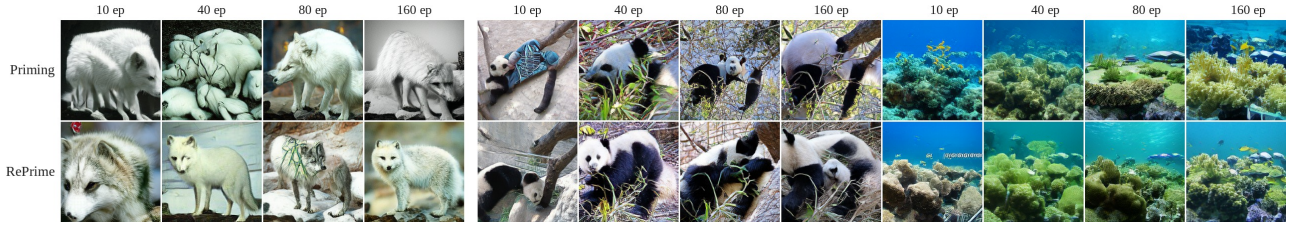


图 S6。**不同预热预算下使用相同噪声**（B，10–160 个预热轮次，每个预热轮次后进行 25k 次后训练更新；上方为 Priming，下方为 RePrime）。预算较低时差异最大：在 10 和 40 个训练轮次时，Priming 生成的北极狐形态扭曲，而 RePrime 已能生成形态完整的动物。

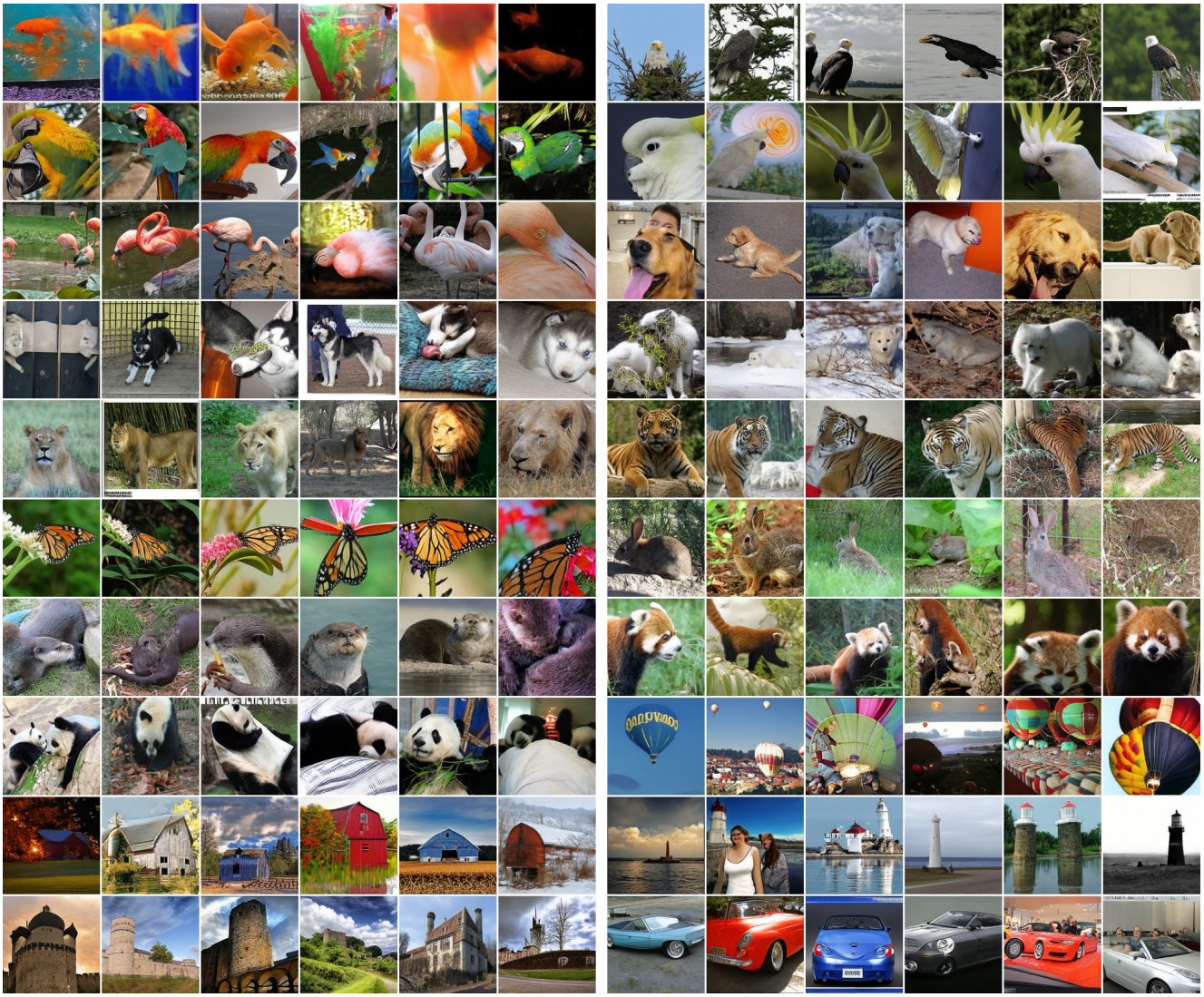


图 S7。**未经筛选的 RePrime-L 样本**（FID 1.155），1 NFE，无引导。每行两个类别，每个类别六个样本。

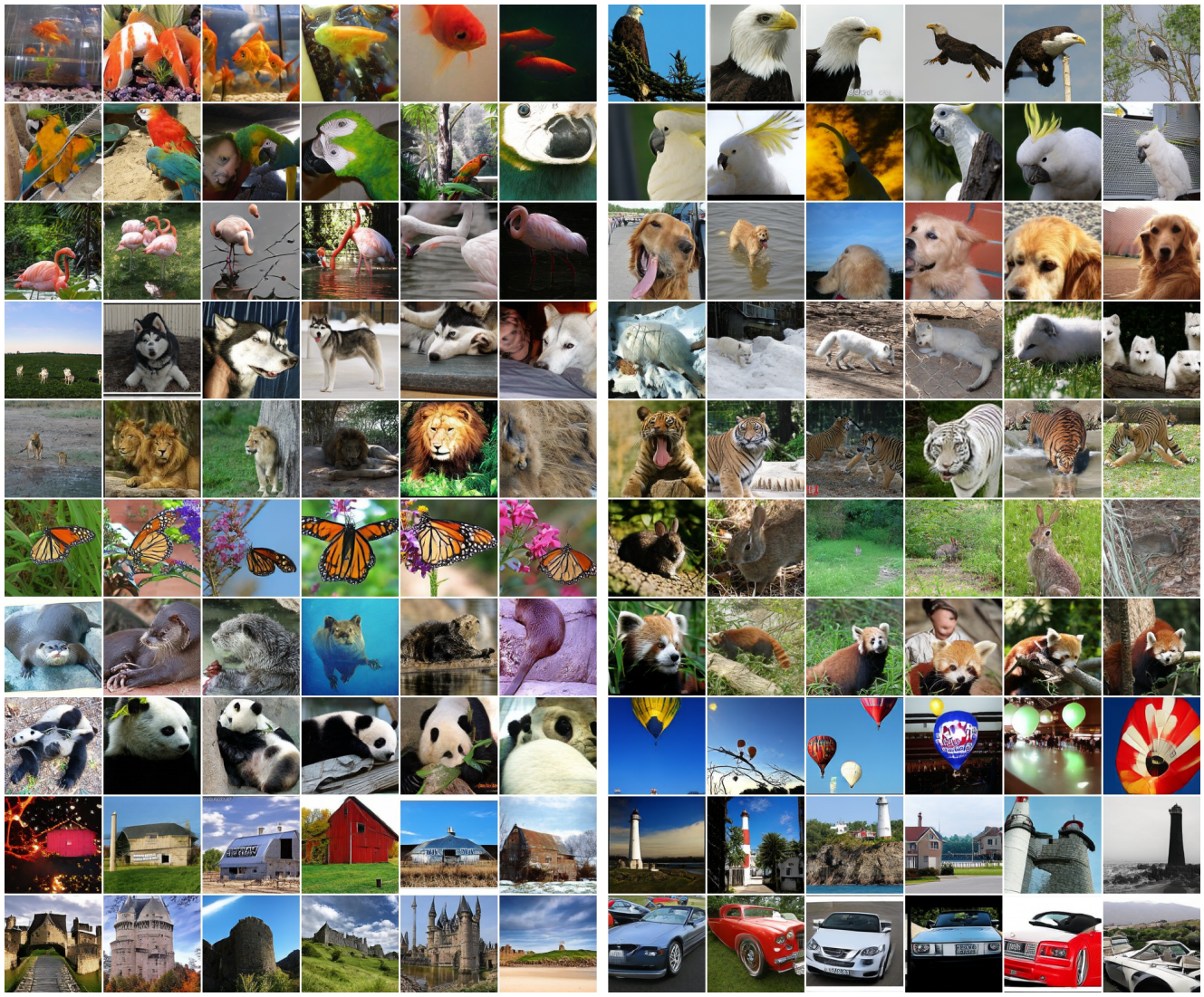


图 S8。**未经筛选的 RePrime-B 样本**（区块 8，200 个预热轮次，FD_r^6 8.15）。

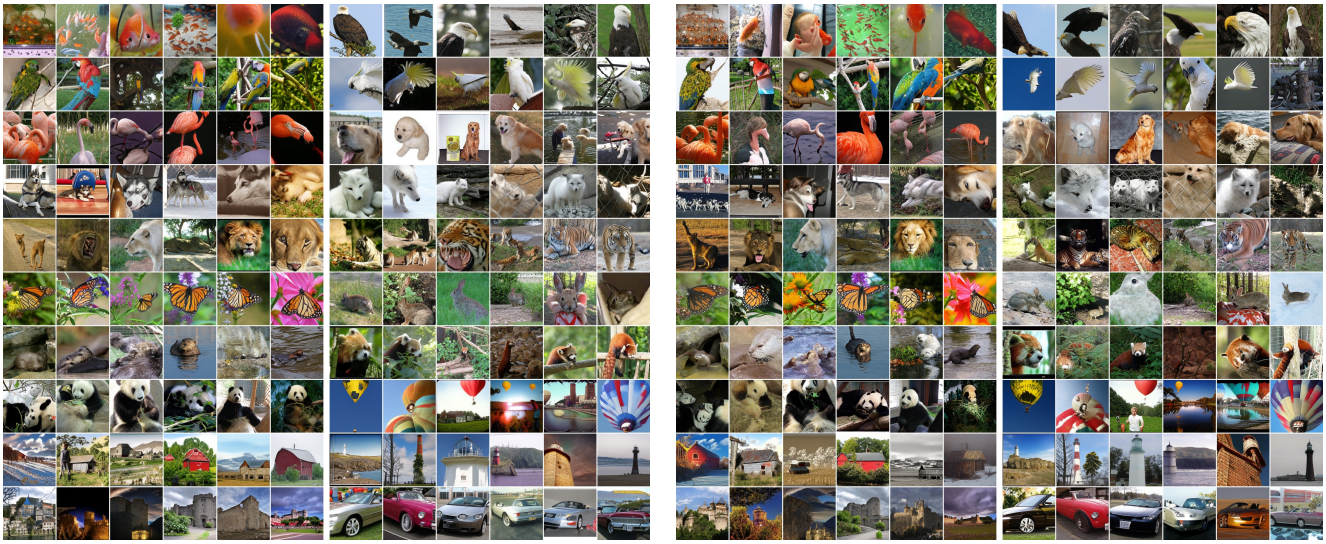


图 S9. **未经筛选的 RePrime-J-B 样本**：经过 240k 次联合更新但未经后训练的原生模型（左；FID 2.22，FD_r^5 9.33），以及经过 25k 次精炼更新步数后的模型（右；FID 1.224，三个随机种子的均值）。