扩散蒸馏的对抗分布匹配
面向高效图像和视频合成

Yanzuo Lu[1,2], Yuxi Ren[2], Xin Xia[2], Shanchuan Lin[2], Xing Wang[2],
Xuefeng Xiao[2*], Andy J. Ma[1,3,4†], Xiaohua Xie[1,3,4,5], Jian-Huang Lai[1,3,4,5]

[1]Sun Yat-Sen University [2]ByteDance Seed Vision
[3]Guangdong Provincial Key Laboratory of Information Security Technology, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[5]Pazhou Lab (HuangPu), Guangzhou, China

oliveryanzuolu@gmail.com,xiaoxuefeng.ailab@bytedance.com,majh8@mail.sysu.edu.cn

Abstract

分布匹配蒸馏 (DMD) 是一种有前途的分数蒸馏技术，它将预训练的教师扩散模型压缩为高效的一步或多步学生生成器。然而，它对反向 Kullback-Leibler(KL) 散度最小化的依赖可能在某些应用中引起模式崩溃 (或模式寻找)。为了规避这一固有缺陷，我们提出了对抗分布匹配 (ADM)，一种新颖的框架，该框架利用基于扩散的判别器在对抗方式中对齐真实和虚假分数估计器之间的潜在预测。在极具挑战性的一步蒸馏的背景下，我们通过在潜在和像素空间中使用混合判别器的对抗蒸馏进一步改进了预训练生成器。有别于在 DMD2 预训练中使用的均方误差，我们的方法结合了从教师模型收集的 ODE 对上的分布损失，从而为下一阶段的分数蒸馏微调提供了更好的初始化。通过将对抗蒸馏预训练与 ADM 微调结合到一个统一的流程中，称为 DMDX，我们提出的方法在 SDXL 上的一步性能优于 DMD2，同时消耗更少的 GPU 时间。额外的实验在 SD3-Medium、SD3.5-Large 和 CogVideoX 上应用多步 ADM 蒸馏，树立了高效图像和视频合成的新基准。

1. 介绍

最近加速扩散模型 [12, 17, 63, 65] 的方法主要集中在通过蒸馏减少采样步骤。蒸馏过程通常训练一个更高效的生成器（也称为学生模型），以近似预训练教师的输出分布。该领域出现了多条路径并行发展，包括渐进蒸馏 [23, 58] 、一致性蒸馏 [18, 34, 35, 43, 55, 66, 69, 70, 92] 、得分蒸馏 [3, 16, 21, 38, 55, 59, 61, 84–87, 93] 、修正流 [27, 28, 30, 71, 81] 和对抗蒸馏 [23, 24, 43, 55, 60, 61, 85, 91] 。它们目前作为明显但非独占的研究方向出现。

分布匹配蒸馏 (DMD) [85, 86] 是一种很有前景的评分蒸馏 [51] 方法，它将强大的文本到图像扩散模型

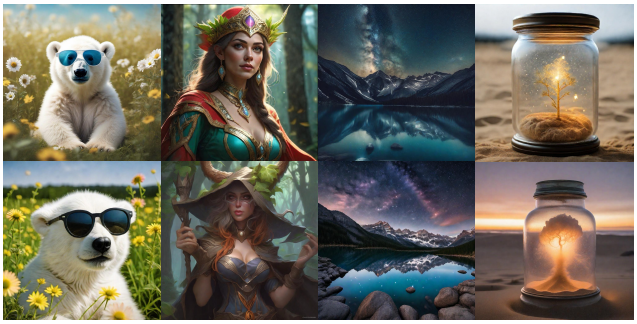

Figure 1. 在这些图像中，有些由基线 SDXL 通过 50NFE 生成，其他则通过我们的 DMDX 在 1NFE 生成。你能分辨出哪个是加速生成的吗？答案在脚注‡ 中。

SDXL-Base [56] 蒸馏成一个具有高保真度的一步生成器。虽然 DMD 的主要贡献在于引入额外的正则项来约束分布匹配损失，但其前身 Diff-Instruct [37] 开创性地使用了虚假评分估计器来近似学生模型的输出分布。这与对抗扩散蒸馏（ADD） [61] 中的蒸馏损失形成对比，后者直接使用学生模型本身作为评分估计器。理论上，可以在文本到三维生成中找到 DMD 和 ADD 中蒸馏损失之间的固有对应性，变分评分蒸馏 (VSD) [74] 和 分数蒸馏采样 (SDS) [51] 之间，其中 SDS 是通过使用单点 Dirac 分布作为变分分布 [74] 的 VSD 的一种特化实例。因此，DMD 中的 VSD 损失比 ADD 中的 SDS 损失表现更好。从直觉上讲，由于扩散模型在少步生成中被蒸馏时能力显著下降，学生模型不再能像教师那样充当一个好的评分估计器。因此，ADD 中基于评分的蒸馏损失几乎不对其最终性能作出贡献。

然而，DMD 损失的优化依赖于反向 库尔贝克-莱布勒（KL）散度最小化，这是一个强制将低概率区域驱动为零的过程，这使得模型只关注几个占主导地位的模式，并可能导致模式崩溃 [45] 。为了增强样本多样性，DMD [86] 使用合成数据的 基于常微分

*Project Lead. † Corresponding Author.
‡ Our DMDX (left to right): bottom, top, bottom, top.

方程 (Ordinary Differential Equation) 正则化器，而 DMD2 [85] 引入了一个基于真实数据的 基于 GAN（生成对抗网络 [8] ）正则化器来抵消这一副作用。后续的努力如 矩匹配蒸馏 (MMD) [59] ，得分身份蒸馏 (SiD) [93] 和 分数隐式匹配（SIM） [38] 都使用 Fisher 散度的变体来使假分数估计器与预训练的真实分数估计器对齐 [16] 。尽管取得了巨大成功，它们在分布匹配方面的能力受到对预定义显式散度度量形式的依赖的限制。在这项工作中，我们想提出并研究一个问题：Can we bypass the limitations of a predefined divergence by developing a framework that learns an implicit, data-driven discrepancy measure, thereby enabling more flexible and fine-grained matching of complex, high-dimensional distributions? 由于复杂多模态的文本条件图像甚至视频生成中的多方面对齐要求可能无法完全捕捉到，这激发了我们探索更具适应性的差异学习范式的动机。作为我们的 first contribution ，我们展示了如何在教师模型的先验知识和动态可学习参数的帮助下，以对抗的方式对真实和伪得分估计器之间的潜在预测进行对齐，从而实现得分蒸馏，这被称为对抗分布匹配（ADM）。这不同于 DMD 和 DMD2 中用于抵消反向 KL 散度的模式崩溃效应的基于 ODE 或 GAN 的正则化器。相反，我们通过 GAN 训练进行分布匹配以替代并规避使用 DMD 损失，我们将在 Sec. 4.1.3 中提供更多讨论。

我们的 second contribution 关注的是极具挑战性的一步蒸馏，我们观察到评分蒸馏存在更高的梯度爆炸和消失的风险。我们将问题更多地归结于学生和教师分布之间的支持集重叠较少，而不仅仅是由于在 DMD2 [85] 中归因于伪评分估计器中的近似误差。换句话说，虽然评分蒸馏能产生优质的生成质量，但它对初始化提出了更高的要求，特别是在极少步骤的蒸馏中，我们将在 Sec. 4.3.2 中提供更深入的分析。尽管我们注意到在 DMD2 的 SDXL 一步蒸馏中，使用了一种基于 ODE 的合成数据预训练方法（详见 https://github.com/tianweiy/DMD2/blob/main/experiments/sdxl/README.md#1-step-sample-trainingtesting-commands-work-in-progress: :text=）。我们的实验表明，当为分布匹配提供更好的初始化时，双时间尺度更新规则（TTUR）[2] 对最终性能的影响非常有限。

自然地，我们的 third contribution 着重于为进一步基于得分的微调提供更好的初始化。我们采用对抗蒸馏技术，借鉴了 SDXL-Lightning [23] 和 LADD [60] 的启发，以在样本质量和模式收敛之间进行权衡。通过这种分布级别的损失优化，我们可以预训练学生模型以捕捉教师模型分布中的更多潜在模式，特别是在我们将于 Sec. 4.2 中介绍的通过潜在空间和像素空间的混合判别器方面。为促进多样性，我们还建议对生成器采用立方时间步长排程，以偏向于更高的噪声水平。

通过将对抗蒸馏预训练与 ADM 微调相结合，形成一个统一的管道，即 DMDX，我们的一步 SDXL 在功能评估次数（NFE）上提供了与基线相比具有竞争力的保真度，并在 × 加速了 50 倍，如 Fig. 1 所示。对包括 SD3-Medium、SD3.5-Large [6] 和 CogVideoX [83] 在内的现有最佳扩散模型进行了多步 ADM 蒸馏的更多实验，一直达到了高效图像和视频合成的新基准。

2. 相关工作

渐进蒸馏在 [58] 中被提出，用于将多步预测提炼为沿着轨迹相同距离的一步预测。SDXL-Lightning [23] 扩展了这个想法，通过使用 GAN 训练首次实现了一步高分辨率（1024px）生成。这种涉及多阶段的过程可能相当繁琐，因为它需要从其前身中反复提炼，每次将采样步骤减半。

一致性蒸馏在 [7, 34, 35, 64, 66] 中提出，将一致性属性引入扩散模型，即针对原始样本的预测对于属于相同轨迹的任意一对噪声时间步保持一致。后续的努力将其扩展为轨迹一致性以放松训练目标，即对于任意后续时间步的噪声样本的预测保持一致，包括 CTM [18] ，TCD [92] ，TSCD [55] 和 PCM [69] 。

位流校正 [27, 28] 旨在通过多次返流过程获得更快的直线路径，这些过程反复学习其前身的众多 ODE 对的速度。PeRFlow [81] 尝试将轨迹拆分成片段，并对真实数据样本应用逐段校正。在本文中，我们凭经验发现，直线性也可以通过对抗性蒸馏范式来满足，而无需重复收集大量合成数据。分数蒸馏直观上试图保持学生模型分布中在特定噪声水平出现的样本，与在教师模型分布中出现的概率相同。基于单一发散可能存在问题的动机，一个并行的工作分数混合蒸馏（SMD） [16] 共享我们的观点，并明确设计了一类 α -偏向 Jensen–Shannon (JS) 发散用于优化。相比之下，我们通过对抗性方式隐式地衡量假得分与真实得分估计器之间的差异，从而促进一个更加有能力和适应性强的差异学习范式，并支持蒸馏过程。对抗蒸馏采用一个鉴别器来将学生模型分布与特定目标分布对齐 [40, 94] 。虽然方法 UFOGen [79] 、DMD2 [85] 和 APT [24] 直接与真实数据分布对齐，SDXL-Lightning [23] 和 Hyper-SD [55] 利用教师模型的中间时间步预测作为近似目标。受到 LADD [60] 的启发，我们的对抗蒸馏预训练与从教师模型生成的基于 ODE 的合成数据对齐。

人类反馈学习最初是由 ReFL [78] 在扩散模型中考虑的，该模型包括奖励模型训练和偏好微调两个阶段。Hyper-SD [55] 将其视为一个独立技术，最终应用 LoRA 插入以将低步生成器的输出分布转向人类偏好。后续研究，尤其是 [36, 39] ，进一步尝试整合 CFG 和基于分数的散度。

3. 预备知识

3.1. 扩散模型

给定一个具有标准差 σ_{data} 的训练数据分布 p_{data} ，扩散模型 [12] 通过逆向扩散过程生成样本，该过程逐步向数据样本 $\boldsymbol{x}_0 \sim p_{data}$ 添加噪声，公式为

$$\boldsymbol{q}(\boldsymbol{x}_t|\boldsymbol{x}) \sim \mathcal{N}(\alpha_t \boldsymbol{x}, \sigma_t^2 \boldsymbol{I}), \tag{1}$$

，其中 $\alpha_t \geq 0, \sigma_t > 0$ 是指定的噪声计划，使得 α_t/σ_t 相对于 t 单调递减，并且较大的 t 表示更大的噪声。我们考虑去噪模型的两种不同表述方式。

DDPM 和 DDIM [12, 63] 假设离散时间安排为 $t \in [1, T]$（通常是 $T = 1000$）和噪声预测参数化 * [58]。训练目标由以下公式给出，

$$\mathbb{E}_{\boldsymbol{x}_0, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}[w(t)\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \epsilon\|_2^2], \tag{2}$$

，其中 $w(t)$ 是加权函数，$\boldsymbol{\epsilon}_\theta$ 是具有参数 θ 的神经网络。噪声安排定义为 $\alpha_t = \sqrt{\bar{\alpha}_t}, \sigma_t = \sqrt{1 - \bar{\alpha}_t}$，使得 $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$。对于 DDIM 采样，它通过 $d\bar{\boldsymbol{x}}_t = \boldsymbol{\epsilon}_\theta(\frac{\bar{\boldsymbol{x}}_t}{\sqrt{\bar{\sigma}_t^2 + 1}})d\bar{\sigma}_t$ 解决概率流常微分方程（PF-ODE）[65]，其中 $\bar{\boldsymbol{x}}_t = \frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}}$ 和 $\bar{\sigma}_t = \sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}}$，从 $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 开始并在 $\boldsymbol{x}_0$ 停止。

流量匹配 [26, 27, 30] 使用速度预测参数化 [58] 和连续时间系数（通常为 $\alpha_t = 1 - t, \sigma_t = t$ 和 $t \in [0, T = 1]$）。条件概率路径或我们所说的速度可以表示为 $\boldsymbol{v}_t = \frac{d\alpha_t}{dt}\boldsymbol{x}_0 + \frac{d\sigma_t}{dt}\epsilon$，因此训练目标是，

$$\mathbb{E}_{\boldsymbol{x}_0, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}[w(t)\|\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{v}_t\|_2^2], \tag{3}$$

其中，$w(t)$ 是一个加权函数，$\boldsymbol{v}_\theta$ 是由 θ 参数化的神经网络。采样过程从 $t = T$ 开始使用 $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$，并在 $t = 0$ 停止，通过 $d\boldsymbol{x}_t = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t)dt$ 解 PF-ODE。

3.2. 分布匹配蒸馏

DMD [85, 86] 通过最小化目标分布 p_{real} 与高效生成器输出分布 p_{fake} 之间的反向 KL 散度，将预训练的扩散模型 $\boldsymbol{F}_\phi(\boldsymbol{x}_t, t)$ 提炼为一步或多步的高效生成器 $\boldsymbol{G}_\theta(\boldsymbol{x}_t, t)$。DMD 目标关于 θ 的梯度为：

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathop{\mathbb{E}}_{\boldsymbol{z}, t', t, \boldsymbol{x}_t} -[(s_{\text{real}}(\boldsymbol{x}_t) - s_{\text{fake}}(\boldsymbol{x}_t))\frac{d\boldsymbol{G}_\theta(\boldsymbol{z}, t')}{d\theta}], \tag{4}$$

其中 $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$，$t'$ 是从预定义的生成器时间表中随机选择的，$t \sim \mathcal{U}(0, T)$，而带噪样本 $\boldsymbol{x}_t = \boldsymbol{q}(\boldsymbol{x}_t|\hat{\boldsymbol{x}}_0)$ 是通过随机扩散生成器输出 $\hat{\boldsymbol{x}}_0 = \boldsymbol{G}_\theta(\boldsymbol{z}, t')$ 获得的。分数函数 $s_{\text{real}}(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t} \log p_{\text{real}}(\boldsymbol{x}_t)$，$s_{\text{fake}}(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t} \log p_{\text{fake}}(\boldsymbol{x}_t)$ 是指向在给定噪声水平 [17, 65] 下数据更高密度的矢量场，分别针对 p_{real} 和 p_{fake}。

虽然真实得分估计器是教师模型 $\boldsymbol{F}_\phi(\boldsymbol{x}_t, t)$ 本身，假得分估计器 $\boldsymbol{f}_\psi(\boldsymbol{x}_t, t)$ 被初始化为和 $\boldsymbol{F}_\phi(\boldsymbol{x}_t, t)$ 相同，并通过与预训练损失 Eqs. (2) and (3) 一起动态学习来描述 p_{fake}。实际上，Eq. (4) 中的梯度计算为，

$$grad(\hat{\boldsymbol{x}}_0, \boldsymbol{x}_t, t) = \frac{\boldsymbol{f}_\psi(\boldsymbol{x}_t, t) - \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)}{\|\hat{\boldsymbol{x}}_0 - \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)\|_1} \tag{5}$$

使得训练损失实现为，

$$\mathcal{L}_{\text{DMD}}(\theta) = \mathop{\mathbb{E}}_{\boldsymbol{z}, t', t, \boldsymbol{x}_t}[\|\hat{\boldsymbol{x}}_0 - sg(\hat{\boldsymbol{x}}_0 - grad(\hat{\boldsymbol{x}}_0, \boldsymbol{x}_t, t))\|_2^2], \tag{6}$$

其中 $sg(\cdot)$ 表示停止梯度操作。

*对于像 CogVideoX [83] 这样的作品来说，这是不强制的，该作品同样使用了 DDIM 采样，但在速度上进行了参数化。不同的参数化可以相互等价地转换为对方 [58]。

4. 方法论

4.1. 对抗性分布匹配

我们没有使用预定义的虚假和真实分布之间的发散度，而是通过对抗性的判别器使用一种隐式的、数据驱动的不一致度量。具体来说，我们的判别器 $\boldsymbol{D}_\tau(\boldsymbol{x}_t, t)$ 由一个冻结的潜在扩散模型组成，该模型的初始化与教师模型 $\boldsymbol{F}_\phi(\boldsymbol{x}_t, t)$ 相同，并在不同的 UNet [57] 或 DiT [49] 块上增加多个可训练的头。给定由少步生成器 $\hat{\boldsymbol{x}}_0 = \boldsymbol{G}_\theta(\boldsymbol{z}, t')$ 的输出扩散而来的有噪声样本 $\boldsymbol{x}_t = \boldsymbol{q}(\boldsymbol{x}_t|\hat{\boldsymbol{x}}_0)$，得分估计器不再针对终点 $\boldsymbol{x}_0^{\text{fake}} = \boldsymbol{f}_\psi(\boldsymbol{x}_t, t)$ 和 $\boldsymbol{x}_0^{\text{real}} = \boldsymbol{F}_\phi(\boldsymbol{x}_t, t)$ 解 PF-ODE，就像在 Eq. (5) 中使用的那样。

相反，我们设置一个固定的时间步长间隔 Δt（默认为 $T/64$）并求解相对于 $(t - \Delta t)$ 的 PF-ODE，使得可以获得虚假的样本 $\boldsymbol{x}_{t-\Delta t}^{\text{fake}}$ 和真实样本 $\boldsymbol{x}_{t-\Delta t}^{\text{real}}$，以作为分数预测并发送到鉴别器中。鉴别器通过多个可学习的头层次聚合来自冻结骨干层的特征，并动态加权它们，建立了一个自适应的差异度量，该度量利用了扩散先验和数据驱动的可训练动态。我们使用铰链损失 [22] 交替训练生成器 $\boldsymbol{G}_\theta(\boldsymbol{x}_t, t)$ 和鉴别器 $\boldsymbol{D}_\tau(\boldsymbol{x}_t, t)$。这鼓励虚假分数预测 $\boldsymbol{x}_{t-\Delta t}^{\text{fake}}$ 更接近真实分数预测 $\boldsymbol{x}_{t-\Delta t}^{\text{real}}$：

$$\mathcal{L}_{\text{GAN}}(\theta) = \mathop{\mathbb{E}}_{\boldsymbol{x}_{t-\Delta t}^{\text{fake}}}[-\boldsymbol{D}_\tau(\boldsymbol{x}_{t-\Delta t}^{\text{fake}}, t - \Delta t)] \tag{7}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(\tau) = \mathop{\mathbb{E}}_{\boldsymbol{x}_{t-\Delta t}^{\text{fake}}, \boldsymbol{x}_{t-\Delta t}^{\text{real}}}[&\max(0, 1 + \boldsymbol{D}_\tau(\boldsymbol{x}_{t-\Delta t}^{\text{fake}}, t - \Delta t)) \\ &+ \max(0, 1 - \boldsymbol{D}_\tau(\boldsymbol{x}_{t-\Delta t}^{\text{real}}, t - \Delta t))]\end{aligned} \tag{8}$$

结合动态学习的虚假模型，我们在 Appendix A 中阐明了我们的训练过程。整体流程在 Fig. 2 中展示。

4.1.1. 判别器时间步 $(t - \Delta t)$ 的动机

鉴于分数蒸馏的最终目标是使学生和教师模型的随噪声水平变化的概率流完全匹配，在衡量分布之间的差异时必须考虑时间步长信息。这与我们的鉴别器设计相符，该设计使用的是预训练的扩散模型，并且我们与 PF-ODE 一起迈出了一小步，成功地保留了分数预测器的输入时间步长信息。

4.1.2. 数据驱动效应

使用判别器作为分布差异度量的灵活性不仅体现在得分函数的噪声水平上，还体现在蒸馏过程中。随着蒸馏的迭代，模型接触到的数据信息越来越多样化，导致两个分布之间的模式差异发生变化。在训练的早期阶段，差异显著时，需要更全面的评估，而在后期阶段，当差异变得较小时，可能需要更加局部化、精细的优化。换句话说，由于数据量的推动，在不同的训练阶段使用的发散度量可能会有所不同。

4.1.3. 与 DMD 和 DMD2 的关系

为了缓解 DMD 损失中的模式崩溃问题，在 DMD [85] 和 DMD2 [85] 中，分别附加使用了基于 ODE 的正则化和基于 GAN 的正则化进行蒸馏。然而，这两个正则化项并没有从根本上解决如 Fig. 4 (a) 所示由反向 KL 散度引入的模式寻求行为，而是通过在损失之间的权

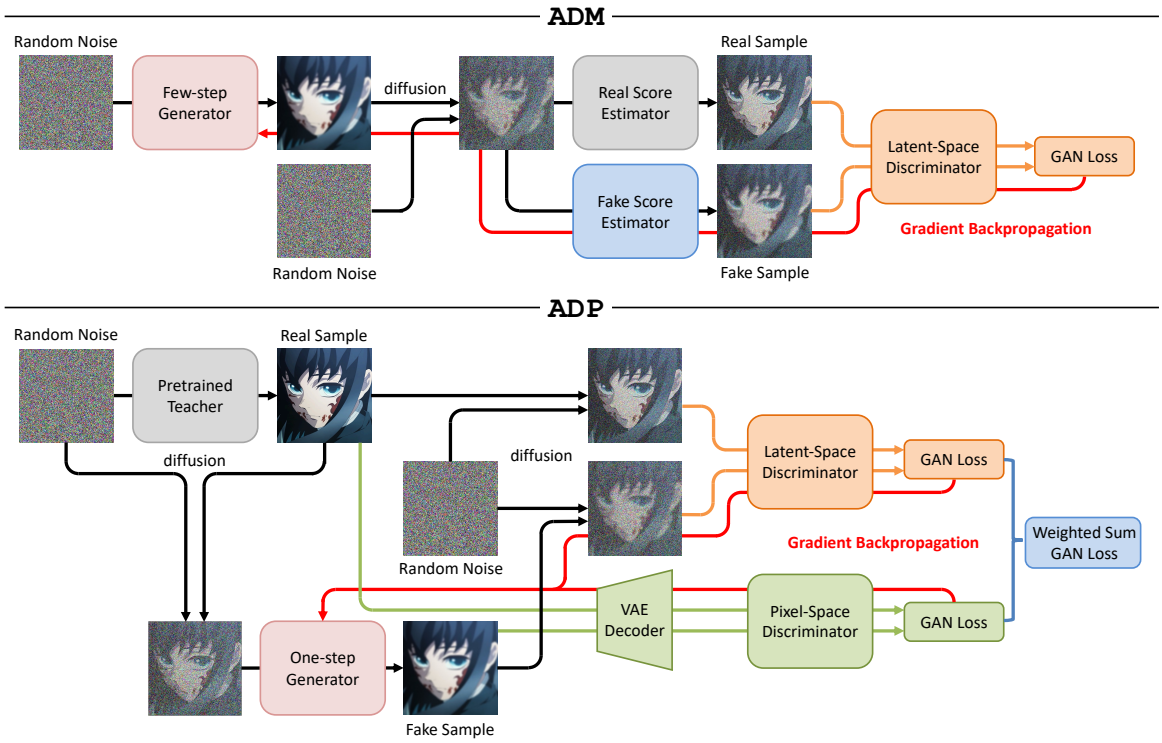

Figure 2. 我们提出的对抗分布匹配（ADM）和对抗蒸馏预训练（ADP）的整体流程。

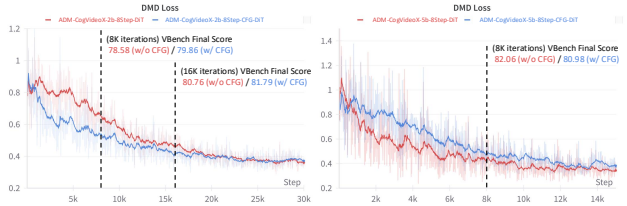

Figure 3. CogVideoX 多步 ADM 蒸馏过程中 DMD 损失的变化。请注意，我们在 ADM 蒸馏过程中并未直接优化这一目标，而是记录它随迭代过程的变化。

衡进行抵消。在 ADM 中，我们的对抗损失实际上扮演了 DMD 损失的角色，通过隐式的、数据驱动的差异度量来实现分数蒸馏，而不是预定义的散度。因此，我们在 ADM 中使用 GAN 训练的动机与 DMD2 [85] 中的不同，我们不需要额外的正则化项。直观上来说，可学习的判别器可以逼近任何非线性函数以隐式测量分布差异度，这可能本质上包含了 DMD 损失中的反向 KL 散度。如 Fig. 3 所示，我们在 CogVideoX [83] 上的多步 ADM 蒸馏过程中可视化了 Eq. (6) 中 DMD 损失的变化。虽然没有直接在 Eq. (6) 上进行优化，但结果显示了一个非常稳定的下降趋势，支持了我们的假设。更多的理论讨论请参阅 Sec. 4.3.3 。

4.2. 对抗蒸馏预训练

为稳定极其困难的单步蒸馏，我们选择通过对合成数据进行对抗蒸馏预训练，为 ADM 微调提供更好的初始化。在多个方面，我们的预训练配置参考了 Rectified Flow [27] ，其中我们 1）以离线方式从教师模型中收集 ODE 对，2）通过在纯噪声和 ODE 对的干净数据样本之间进行线性插值来构建噪声样本，3）将生成器的预测目标更改为 ODE 对的速度。

对于对抗训练，我们分别构建了一个从教师模型初始化的潜在空间判别器 $\boldsymbol{D}_{\tau_1}(\boldsymbol{x}_t, t)$ 和一个从 SAM 的视觉编码器 [19] 模型初始化的像素空间判别器 $\boldsymbol{D}_{\tau_2}(\boldsymbol{x})$ ，如 Fig. 2 所示。我们还在这两个主干网络上附加多个可训练的头部，类似于 ADM 中的做法。所有这些都有助于增加辨别能力，促进学生模型在教师模型分布中发现更多潜在模式。具体来说，令 $\tilde{\boldsymbol{x}}_0 = \boldsymbol{G}_\theta(\boldsymbol{x}_t, t)$ 表示生成器预测的 PF-ODE 终点，其中 $\boldsymbol{x}_t$ 表示在一个随机时间步长 $t \in [0, T]$ 的 ODE 对之间插值的噪声样本。对于潜在空间判别器，我们将生成器输出与另一个随机噪声和时间步长 $t' \in (0, T]$ 进行扩散，得到 $\tilde{\boldsymbol{x}}_{t'} = \boldsymbol{q}(\tilde{\boldsymbol{x}}_{t'}|\tilde{\boldsymbol{x}}_0)$ 作为其输入。对于像素空间判别器，生成器输出将首先通过 VAE 解码器解码，然后送入视觉编码器。训练目标基于 Hinge loss [22] ，鼓励生成器输出 $\tilde{\boldsymbol{x}}_0$ 更加接近合成的数据样本 $\boldsymbol{x}_0$ ：

$$\mathcal{L}_{\text{GAN}}(\theta) = \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}_0, t'} -[\lambda_1 \boldsymbol{D}_{\tau_1}(\tilde{\boldsymbol{x}}_{t'}, t') + \lambda_2 \boldsymbol{D}_{\tau_2}(\tilde{\boldsymbol{x}}_0)] \tag{9}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(\tau_1, \tau_2) = \mathop{\mathbb{E}}_{\boldsymbol{x}_0, \tilde{\boldsymbol{x}}_0, t'} [&\lambda_1 \cdot \max(0, 1 + \boldsymbol{D}_{\tau_1}(\tilde{\boldsymbol{x}}_{t'}, t')) \\ &+\lambda_2 \cdot \max(0, 1 + \boldsymbol{D}_{\tau_2}(\tilde{\boldsymbol{x}}_0)) \\ &+\lambda_1 \cdot \max(0, 1 - \boldsymbol{D}_{\tau_1}(\boldsymbol{x}_{t'}, t')) \\ &+\lambda_2 \cdot \max(0, 1 - \boldsymbol{D}_{\tau_2}(\boldsymbol{x}_0))]\end{aligned} \tag{10}$$

我们通过实验证明，设置平衡系数 $\lambda_1 = 0.85, \lambda_2 = 0.15$ 能够产生视觉上连贯的结果。

4.2.1. 立方生成器时间步调度

直观上，更高的噪声水平通过削弱潜在表示中编码的限制性信息，鼓励探索新的模式。因此，我们建议为生成器采用立方时间步长调度。该调度将均匀的 $[0, T)$ 样本通过 $[1-(t/T)^3]*T$ 映射，非线性地将值集中在 T 附近，伴有类似 LADD [60] 的高噪声。

与 ADM 中首先将判别器用于评分蒸馏不同，潜在空间判别器的使用在对抗蒸馏工作中很常见。受 SDXL-Lightning 的启发，他们发现扩散编码器在较低时间步专注于高频细节，而在较高时间步则专注于低频结构，我们为判别器时间步设置了一个均匀的 $(0, T]$ ，以在预训练过程中捕捉这两种优势。

我们在合成数据上进行对抗蒸馏的动机来源于 LADD [60] ，但在许多方面有所不同，我们 1）通过类似于 Rectified Flow [27] 的方法，以 ODE 对的方式构建噪声样本，而不是随机噪声，2）开发一个立方生成器时间步进调度，以促进确定性欧拉采样而不是一致性采样，3）引入一个额外的像素空间编码器，以提高判别器的能力并找到更多的模式。

4.3. 讨论

4.3.1. ADM 与 ADP 的区别

一个问题可能看起来像 what is the difference between these two adversarial approaches with latent-space discriminators? ：得分蒸馏的有效性与得分函数 $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}; \sigma(t))$ 在不同噪声水平 $\sigma(t)$ 下的定义有关。相比之下，对抗蒸馏仅在 $t=0$ 时对齐干净数据样本的分布。而预训练中的潜在空间鉴别器通过随机扩散生成器输出捕捉不同尺度和细节的信息，ADM 的关键不仅在于此，而是在于解决两个得分估计器的 PF-ODE。换句话说，ADM 还通过在各自分布的不同噪声水平下具有较高密度的噪声样本来监督完整的去噪过程。

这导致在 ADM 中当两个分布在支持集上初始化的重叠较少时，噪声样本会停留在彼此不熟悉的区域，判别器可以很容易地区分它们，从而导致极端的梯度信号。然而，由于高斯噪声是各向同性的，我们人为地为 ADP 中随机扩散的样本创建重叠区域以使区分更加困难，从而导致相对平滑的梯度。因此，我们的 ADM 仍然属于分数蒸馏，因为它鼓励整个概率流更接近，而预训练属于对抗蒸馏，因为它只关心 $t=0$ 的干净数据分布。

4.3.2. 预训练的重要性

我们尚未讨论的另一个问题是 why do we need pre-training for one-step score distillation? ，以 DMD 损失中使用的反向 KL 散度为例：

$$\mathbb{D}_{\mathrm{KL}}(p_{\mathrm{fake}} \| p_{\mathrm{real}}) = \int p_{\mathrm{fake}}(\boldsymbol{x}) \log \frac{p_{\mathrm{fake}}(\boldsymbol{x})}{p_{\mathrm{real}}(\boldsymbol{x})} d\boldsymbol{x}. \quad (11)$$

当采用单步蒸馏时，与多步采样相比，生成器输出在视觉保真度和结构完整性方面较差，导致在某些区域出现 $p_{\mathrm{fake}}(\boldsymbol{x}) \to 0$ ，其中 $p_{\mathrm{real}}(\boldsymbol{x}) > 0$ 。被积函数 $p_{\mathrm{fake}}(\boldsymbol{x}) \log \frac{p_{\mathrm{fake}}(\boldsymbol{x})}{p_{\mathrm{real}}(\boldsymbol{x})}$ 接近于零 $0 \cdot (-\infty)$ 导致优化避开那些 $p_{\mathrm{real}}(\boldsymbol{x}) > 0$ 但 $p_{\mathrm{fake}}(\boldsymbol{x})$ 密度可以忽略的区域，这一现象称为零强制（zero-forcing）。p_{fake} 没有完全覆盖 p_{real} 的支持，而是收敛到 p_{real} 的子集模式，导致如 Fig. 4 （a）所示的模式追求行为。在训练过程中，这有时表现为梯度消失。相反，扩散模型通常通过一步产生的模糊样本也不在教师模型分布中，导致在某些区域出现 $p_{\mathrm{real}}(\boldsymbol{x}) \to 0$ ，其中 $p_{\mathrm{fake}}(\boldsymbol{x}) > 0$ ，并且被积函数 $p_{\mathrm{fake}}(\boldsymbol{x}) \log \frac{p_{\mathrm{fake}}(\boldsymbol{x})}{0}$ 发散至 $+\infty$ ，从而导致数值不稳定和梯度爆炸。

类似地，当学生和教师分布的支持集几乎不重叠时，前向 KL 散度趋近于 $+\infty$ ，其中 $p_{\mathrm{fake}}(\boldsymbol{x}) > 0$ ，而 JS 散度饱和到一个常数 $\log 2$ ，而 Fisher 散度可能在没有定义的情况下退化。因此，当这个假设被破坏时，许多单一散度不再适用，更好的初始化与更多重叠区域变得至关重要，如 Fig. 4 (b) 中所示。

4.3.3. 理论目标

最后一个问题是 why ADM is better than DMD loss theoretically? 事实上，我们使用的 Hinge GAN [22] 已被证明可以最小化总变差距离（TVD） [68] ：

$$TV(p_{\mathrm{fake}}, p_{\mathrm{real}}) = \int |p_{\mathrm{fake}}(\boldsymbol{x}) - p_{\mathrm{real}}(\boldsymbol{x})| d\boldsymbol{x} \quad (12)$$

也就是说，当判别器足够丰富且经过良好训练时，Hinge 损失在收敛时的理论最优值最小化 TVD。当假分布和真实分布的支持集重叠最小时，TVD 相对于反向 KL 散度提供两个关键优势：1）对称性：无论初始分布如何，TVD 都提供相同的差异测度，而不对称的反向 KL 可能表现为模式寻找行为，并忽略整体分布的其他部分。例如，如 Fig. 4 （c）所示，在 $p_{\mathrm{fake}}(\boldsymbol{x}) \to 0$ 时，TVD 保持显著的损失值，并提供覆盖模式的优化方向，而 $p_{\mathrm{real}}(\boldsymbol{x}) > 0$ ，反向 KL 散度在这种情况下因为梯度消失问题而表现不佳，如 Sec. 4.3.2 所讨论。2）有界性：TVD 有界在 [0,1] 之间，因此在训练过程中减轻异常数值的干扰，特别是在我们高维多模态文本条件的图像和视频分布中，避免了反向 KL 散度中由梯度爆炸引起的数值不稳定问题。

Method	Step	NFE	CLIP Score	Pick Score	HPSv2	MPS
ADD [61] (512px)	1	1	35.0088	22.1524	27.0971	10.4340
LCM [34]	1	2	28.4669	20.1267	23.8246	4.8134
Lightning [23]	1	1	33.4985	21.9194	27.1557	10.2285
DMD2 [85]	1	1	35.2153	22.0978	27.4523	10.6947
DMDX (Ours)	1	1	35.2557	22.2736	27.7046	11.1978
SDXL-Base [56]	25	50	35.0309	22.2494	27.3743	10.7042

Table 1. 对 SDXL-Base 进行完全微调的定量结果。

5. 实验

模型。在单步蒸馏中，我们采用对抗性蒸馏预训练（ADP）和在 SDXL-Base [56] 上进行 ADM 微调，称为 DMDX 。对于多步蒸馏，我们仅在文本到图像模

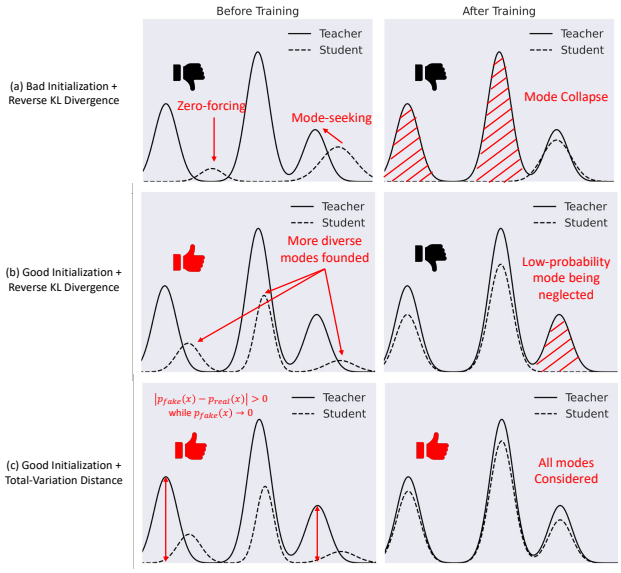

Figure 4. 理论讨论的例证。

Method	Step	NFE	CLIP Score	Pick Score	HPSv2	MPS
TSCD [61]	4	8	34.0185	21.9665	27.2728	10.8600
PCM [69] (Shift=1)	4	4	33.5042	21.9703	27.3680	10.5707
PCM [69] (Shift=3)	4	4	33.3818	21.9396	27.1146	10.5635
PCM [69] (Stoch.)	4	4	33.4185	21.8822	27.3177	10.5200
Flash [3]	4	4	34.3978	22.0904	27.2586	10.6634
ADM (Ours)	4	4	34.9076	22.5471	28.4492	11.9543
SD3-中等 [6]	25	50	34.7633	22.2961	27.9733	11.3652
LADD [60]	4	4	34.7395	22.3958	27.4923	11.4372
ADM (Ours)	4	4	34.9730	22.8842	27.7331	12.2350
SD3.5-大 [6]	25	50	34.9668	22.5087	27.9688	11.5826

Table 2. 关于 LoRA 微调 SD3-Medium 和完全微调 SD3.5-Large 的量化结果。

Method	Step	NFE	Final Score	Quality Score	Semantic Score
ADM	8	8	78.584	80.825	69.621
+ Longer Training × 2	8	8	80.764	83.031	71.693
ADM w/ CFG	8	16	79.865	80.938	75.569
+ Longer Training × 2	8	16	81.796	83.008	76.947
CogVideoX-2b [83]	100	200	80.036	80.801	76.974
ADM	8	8	82.067	83.227	77.423
ADM w/ CFG	8	16	80.982	82.165	76.251
CogVideoX-5b [83]	100	200	81.226	81.785	78.987

Table 3. 完全微调 CogVideoX 的定量结果。

型 SD3-Medium, SD3.5-Large [6] 和文本到视频模型 CogVideoX-2b, CogVideoX-5b [83] 上进行 ADM 训练。根据多数同时进行的工作，我们在文本到图像模型中没有使用无分类器指导（CFG） [11] 。我们尝试在文本到视频模型上进行 CFG 整合实验，方法详见 Sec. 5.2 。

数据集。本研究中提出的 ADP 和 ADM 均无需视觉数据。对于图像生成器，我们使用 JourneyDB [67] 中的文本提示进行训练，该提示具有高度的细节和特

Figure 5. 对 SDXL-Base 进行全量微调的定性结果。

异性。对于视频生成器，我们从 OpenVid-1M [47] 、Vript [82] 和 Open-Sora-Plan-v1.1.0 [50] 收集训练提示。

评价。图像生成器在来自 COCO 2014 [25] 的 1 万个提示上按照 DMD2 [85] 进行评估。我们报告了 CLIP 分数 [52] 和人类偏好基准 PickScore [20] 、HPSv2 [75] 和 MPS [90] ，如许多同时工作的 Hyper-SD [55] 和 Emu3 [73] 。但我们没有在一步定量比较中包括 Hyper-SD ，因为一步 Hyper-SDXL 已直接通过 ReFL [78] 针对人类反馈进行了优化。相反，我们与其中提出的 SD3-Medium [13] 上的 TSCD 算法进行比较，因为 4 步 Hyper-SD3 LoRA 未经过 ReFL 优化。视频生成器在 VBench [14] 上进行评估，该评估在多个质量和语义维度上全面进行。

超参数。尽管在我们提出的 ADP 和 ADM 中需要训练多个模型，但我们在没有进行大量超参数调整的情况下，仍然达到了令人满意的视觉逼真度和结构完整性。在本文的其余部分中，我们仅对生成器的学习率进行调整以适应不同的实验。对于判别器和伪模型的优化器设置在所有实验中都是统一的。除非特别注明需要更长的训练时间，否则我们对于文本到图像和文本到视频模型，仅以批量大小分别为 128 和 8 进行 8K 次迭代的训练。更具体的实现细节在 Appendix B 中提供。

Ablation	CLIP Score	Pick Score	HPSv2	MPS
Ablation on adversarial distillation.				
A1: Rectified Flow [27]	27.4376	20.0211	23.6093	4.4518
A2: DINOv2 as pixel-space	34.1836	21.8750	27.1039	10.2407
A3: $\lambda_1 = 0.7, \lambda_2 = 0.3$	33.6943	21.6344	26.8902	9.9633
A4: $\lambda_1 = 1.0, \lambda_2 = 0.0$	33.8929	21.7395	26.7869	10.0757
A5: w/o ADM (ADP only)	35.7723	22.0095	27.3499	10.6646
Ablation on score distillation.				
B1: ADM w/o ADP	32.5020	21.7631	26.8732	10.8986
B2: DMD Loss w/o ADP	32.7482	21.0341	25.9680	8.8977
B3: DMD Loss w/ ADP	34.5119	21.9366	27.3985	10.6046
B4: DMDX (Ours)	35.2557	22.2736	27.7046	11.1978

Table 4. 消融研究的定量结果。

5.1. 高效图像合成

Tab. 1 定量比较了我们结合 ADP 和单步 ADM 蒸馏的两阶段方法与现有单步蒸馏方法在完全微调 SDXL-Base 上的表现。结果表明，我们的方法在图文对齐和人类偏好方面都取得了卓越的性能，这与 Fig. 5 中的定性比较一致，包括更好的肖像美感、动物毛发细节、主体与背景分离以及物理结构。

对于多步 ADM 蒸馏，它可以作为一种独立的评分蒸馏方法。我们尝试了完全微调和 LoRA 微调 [13] 配置，Tab. 2 中的定量结果证明了我们卓越的性能。Appendix C 中提供了定性结果。

5.2. 高效视频合成

正如 Tab. 3 中定量显示的那样，除了对两种 CogVideoX 尺寸进行常规的 8 步 ADM 蒸馏，我们还尝试将无分类器引导（CFG） [11] 集成到文本到视频任务中。具体来说，我们通过在 $[5.0, 7.0]$ 范围内随机采样来为真实模型分配 CFG 比例，同时通过从真实模型的值中明确减去 2.0 来分配少步生成器的比例。根据经验，我们发现这种减去的偏移需要随着目标采样步数的减少而逐渐增大。虚假模型没有 CFG。VBench [14] 的结果表明，我们的少步生成器在与底层模型进行比较时实现了 92-96 % 的加速。受到 Fig. 3 表明在 8K 次迭代时 DMD 损失没有很好收敛的观察启发，我们对 2B 模型进行了更长时间的训练进行额外评估。实验结果表明，在 ADM 蒸馏过程中，可学习的判别器也对 DMD 损失进行了近似优化。更多定量和定性结果参见 Appendix C 。

5.3. 消融研究

在 Tab. 4 中，我们对全面微调 SDXL-Base 进行了广泛的消融研究以验证我们的有效性。定性比较在 Appendix C 中提供。

ADP 的效果。可以得出以下结论：1）单次回流过程提供的效果非常有限（ A1 ），而多次过程实现有效整改需要相当大的计算开销。2）使用 SAM [19] 比广泛采用的 DINOv2 [48] （ A2/A5 ）提供更多的成像保真性，这可能是因为 SAM 的分辨率为 1024px（相比于 DINOv2 的 518px）更高。3）像素空间的加权 λ_2 不能太大或太小。过大的加权会导致结构完整性下降（ A3/A5 ），而加权不足则导致模糊（ A4/A5 ）。Appendix C 中的定性结果提供了更明显的区别。

ADM 的效果。我们总结了关键发现：1) 缺乏 ADP 会导致显著的性能下降（ B1/B4 ），这与我们在 Sec. 4.3.2 中的分析一致。2) 如果没有正则化器，DMD 损失相比于独立的 ADM 表现不佳（ B1/B2 ），表明其鲁棒性差。3) 虽然 DMD 损失优化也受益于 ADP （ B2/B3 ），但其分布匹配能力仍然不如 ADM （ B3/B4 ）。

TTUR 的效果。Tab. 5 展示了不同 TTUR 设置对最终性能和训练时间的影响。结果显示，增加 TTUR 仅带来微小的性能提升，却几乎翻倍了训练时间，使得这种权衡显然不值得。这突出了我们提出的 ADP 在一步蒸馏中的关键作用，并表明 DMD2 中训练不稳定可能源于支持集重叠不足。

多样性评估。根据 DMD2 [85] ，我们在 Partiprompts [88] 上生成每个提示 4 个样本，使用不同的种子，并在 Tab. 6 中报告平均的成对 LPIPS 相似性 [89] 。结果表明，我们的方法在多样性方面显著优于其他方法。更多随机精选的多种种子样本可以在 Appendix C 中找到。

TTUR	Training Time	CLIP Score	Pick Score	HPSv2	MPS
1	× 1.00	35.2557	22.2736	27.7046	11.1978
4	× 1.85	35.2583	22.2773	27.7255	11.2720
8	× 2.53	35.3299	22.2883	27.7586	11.2838

Table 5. 关于双时间尺度更新规则的消融研究。

	ADD	LCM	Lightning	DMD2	Ours	教师
LPIPS ↑	0.6071	0.6257	0.6707	0.6715	0.7156	0.6936

Table 6. 对 PartiPrompts [88] 进行定量多样性评估。

6. 局限性

我们意识到一个弱点是教师模型可能需要 CFG 才能产生准确的分数预测。我们的实验表明，这是分数蒸馏方法的一种普遍特征，而不是我们方法特有的局限。这限制了对 FLUX.1-dev [1] 等指南蒸馏模型的应用，这可能是未来研究的一个有前途的话题。

7.

致谢 这项工作部分得到了中国国家自然科学基金（U22A2095，12326618，62276281）、广东省基础与应用基础研究基金（2024A1515011882）和广东省信息安全技术重点实验室项目（2023B1212060026）的资助。

References

[1] Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024.

[2] Naresh Babu Bynagari. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017.

[3] Clement Chadebec, Onur Tasar, Eyal Benaroche, and Benjamin Aubin. Flash diffusion: Accelerating any conditional diffusion model for few steps image generation. arXiv preprint arXiv:2406.02347, 2024.

[4] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174, 2016.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In ICML, 2024.

[7] Zhengyang Geng, Ashwini Pokle, William Luo, Justin Lin, and J Zico Kolter. Consistency models made easy. In ICLR, 2025.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NeurIPS, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[10] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In NeurIPS Workshops, 2021.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020.

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In ICLR, 2022.

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In CVPR, 2024.

[15] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. arXiv preprint arXiv:2309.14509, 2023.

[16] Tejas Jayashankar, J. Jon Ryu, and Gregory Wornell. Score-of-mixture training: Training one-step generative models made simple via score estimation of mixture distributions. arXiv preprint arXiv:2502.09609, 2025.

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In NeurIPS, 2022.

[18] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In ICLR, 2024.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In ICCV, pages 3992–4003, 2023.

[20] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In NeuriPS, 2023.

[21] Jonas Kohler, Albert Pumarola, Edgar Schönfeld, Artsiom Sanakoyeu, Roshan Sumbaly, Peter Vajda, and Ali Thabet. Imagine flash: Accelerating emu diffusion models with backward distillation. arXiv preprint arXiv:2405.05224, 2024.

[22] Jae Hyun Lim and Jong Chul Ye. Geometric gan. arXiv preprint arXiv:1705.02894, 2017.

[23] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929, 2024.

[24] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, Lu Jiang, and ByteDance Seed. Diffusion adversarial post-training for one-step video generation. arXiv preprint arXiv:2501.08316, 2025.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In ECCV, 2014.

[26] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2023.

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In ICLR, 2022.

[28] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In ICLR, 2024.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR, 2019.

[30] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. arXiv preprint arXiv:2410.11081, 2024.

[31] Yanzuo Lu, Manlin Zhang, Yiqi Lin, Andy J. Ma, Xiaohua Xie, and Jianhuang Lai. Improving pre-trained

masked autoencoder via locality enhancement for person re-identification. In PRCV, pages 509–521, 2022.

[32] Yanzuo Lu, Meng Shen, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Mlnet: Mutual learning network with neighborhood invariance for universal domain adaptation. In AAAI, pages 3900–3908, 2024.

[33] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In CVPR, pages 6420–6429, 2024.

[34] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378, 2023.

[35] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556, 2023.

[36] Weijian Luo. Diff-instruct++: Training one-step text-to-image generator model to align with human preferences. In TMLR, 2024.

[37] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In NeurIPS, pages 76525–76546, 2023.

[38] Weijian Luo, Zemin Huang, Zhengyang Geng, J. Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. In NeurIPS, 2024.

[39] Weijian Luo, Colin Zhang, Debing Zhang, and Zhengyang Geng. David and goliath: Small one-step model beats large diffusion with score post-training. In ICML, 2025.

[40] Yihong Luo, Xiaolong Chen, Xinghua Qu, Tianyang Hu, and Jing Tang. You only sample once: Taming one-step text-to-image synthesis by self-cooperative diffusion gans. In ICLR, 2025.

[41] Hongxu Ma, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In ACMMM, 2025.

[42] Hongxu Ma, Chenbo Zhang, Lu Zhang, Jiaogen Zhou, Jihong Guan, and Shuigeng Zhou. Fine-grained zero-shot object detection. In ACMMM, 2025.

[43] Xiaofeng Mao, Zhengkai Jiang, Fu-Yun Wang, Wenbing Zhu, Jiangning Zhang, Hao Chen, Mingmin Chi, and Yabiao Wang. Osv: One step is enough for high-quality image to video generation. arXiv preprint arXiv:2409.11367, 2024.

[44] Yuxi Mi, Zhizhou Zhong, Yuge Huang, Qiuyang Yuan, Xuan Zhao, Jianqing Xu, Shouhong Ding, Shaoming Wang, Rizen Guo, and Shuigeng Zhou. Data synthesis with diverse styles for face recognition via 3dmm-guided diffusion. In CVPR, pages 21203–21214, 2025.

[45] Thomas Minka. Divergence measures and message passing. Microsoft Research, Technical Report, 2005.

[46] Movie Gen Team. Movie gen: A cast of media foundation models. https://ai.meta.com/static-resource/movie-gen-research-paper, 2024.

[47] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371, 2024.

[48] Maxime Oquab, Timothée Darcet, Théo 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, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. TMLR, 2024.

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In ICCV, 2023.

[50] PKU-Yuan Lab and Tuzhan AI. Open-sora-plan-v1.1.0. https://huggingface.co/datasets/LanguageBind/Open-Sora-Plan-v1.1.0, 2024.

[51] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In ICLR, 2023.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021.

[53] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In SC, 2020.

[54] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. arXiv preprint arXiv:1710.05941, 2017.

[55] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In NeurIPS, 2024.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[58] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In ICLR, 2022.

[59] Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. In NeurIPS, 2024.

[60] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast

high-resolution image synthesis with latent adversarial diffusion distillation. arXiv preprint arXiv:2403.12015, 2024.

[61] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In ECCV, 2024.

[62] Meng Shen, Yanzuo Lu, Yanxu Hu, and Andy J. Ma. Collaborative learning of diverse experts for source-free universal domain adaptation. In ACM MM, pages 2054–2065, 2023.

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In ICLR, 2021.

[64] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In ICLR, 2024.

[65] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In ICLR, 2021.

[66] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In ICML, 2023.

[67] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. arXiv preprint arXiv:2307.00716, 2023.

[68] Zhiqiang Tan, Yunfu Song, and Zhijian Ou. Calibrated adversarial algorithms for generative modelling. Stat, 8:e224, 2019.

[69] Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Hongsheng Li, and Xiaogang Wang. Phased consistency model. arXiv preprint arXiv:2405.18407, 2024.

[70] Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Computation-efficient personalized style video generation without personalized video data. In SIGGRAPH ASIA Technical Communications, 2024.

[71] Fu-Yun Wang, Ling Yang, Zhaoyang Huang, Mengdi Wang, and Hongsheng Li. Rectified diffusion: Straightness is not your need in rectified flow. arXiv preprint arXiv:2410.07303, 2024.

[72] Shibo Wang and Pankaj Kanwar. Bfloat16: The secret to high performance on cloud tpus. https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus, 2019.

[73] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869, 2024.

[74] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In NeurIPS, 2023.

[75] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341, 2023.

[76] Yuxin Wu and Kaiming He. Group normalization. In ECCV, 2018.

[77] Xuefeng Xiao, Lianwen Jin, Yafeng Yang, Weixin Yang, Jun Sun, and Tianhai Chang. Building fast and compact convolutional neural networks for offline handwritten chinese character recognition. Pattern Recognition, 72:72–81, 2017.

[78] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In NeurIPS, 2023.

[79] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In CVPR, 2024.

[80] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In CVPR, pages 15909–15919, 2025.

[81] Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. In NeurIPS, 2024.

[82] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. arXiv preprint arXiv:2406.06040, 2024.

[83] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024.

[84] Hongwei Yi, Shitong Shao, Tian Ye, Jiantong Zhao, Qingyu Yin, Michael Lingelbach, Li Yuan, Yonghong Tian, Enze Xie, and Daquan Zhou. Magic 1-for-1: Generating one minute video clips within one minute. arXiv preprint arXiv:2502.07701, 2025.

[85] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In NeurIPS, 2024.

[86] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In CVPR, pages 6613–6623, 2024.

[87] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. arXiv preprint arXiv:2412.07772, 2024.

[88] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. TMLR, 2022.

[89] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, pages 586–595, 2018.

[90] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In CVPR, pages 8018–8027, 2024.

[91] Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, and Jian Ren. Sf-v: Single forward video generation model. In NeurIPS, 2024.

[92] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping. arXiv preprint arXiv:2402.19159, 2024.

[93] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In ICML, 2024.

[94] Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In ICLR, 2025.

扩散蒸馏的对抗分布匹配面向高效图像和视频合成

Supplementary Material

A. 对抗性分布匹配

在 ADM 蒸馏过程中，虚假分数估计器、生成器和判别器交替更新。下面的 Algorithm 1 阐明了训练过程。我们的 Sec. 5.3 中的消融实验表明，TTUR 对最终性能影响很小。因此，在我们的实验中，我们将 TTUR 设置为 1，这意味着虚假模型和生成器以相同的频率更新。

Algorithm 1 ADM 训练过程

Input: pretrained teacher model as real score estimator $\boldsymbol{F}_\phi$
Output: few-step generator $\boldsymbol{G}_\theta$ with schedule $\{t_0, t_1, ..., t_N\}$
Initialize: fake score estimator $\boldsymbol{f}_\psi \leftarrow \boldsymbol{F}_\phi$, generator $\boldsymbol{G}_\theta \leftarrow \boldsymbol{F}_\phi$, latent-space discriminator $\boldsymbol{D}_\tau \leftarrow \boldsymbol{F}_\phi$ with multiple trainable heads, generator iteration $genIter \leftarrow 0$, global iteration $globalIter \leftarrow 0$
while $genIter < maxIter$ do
  $globalIter$ += 1

  // *update fake score estimator* $\boldsymbol{f}_\psi$
  sample pure noise $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  solve the PF-ODE w.r.t. N steps in schedule $\boldsymbol{x}_0 \leftarrow \boldsymbol{G}_\theta(\boldsymbol{z}, \cdot)$
  sample new pure noise $\boldsymbol{z}_f$ and random timestep t_f
  update ψ with $(\boldsymbol{x}_0, t_f, \boldsymbol{z}_f)$ and pretrain loss in Eq. (2) or Eq. (3)
  if not ($globalIter$ % TTUR) == 0 then continue

  // *update generator* $\boldsymbol{G}_\theta$
  sample pure noise $\hat{\boldsymbol{z}}$ and random index $n \in [1, N]$
  solve the PF-ODE w/o grad following $t_N \rightarrow t_{N-1} \rightarrow ... \rightarrow t_n$, i.e. $\hat{\boldsymbol{z}} = \hat{\boldsymbol{x}}_{t_N} \rightarrow \hat{\boldsymbol{x}}_{t_{N-1}} \rightarrow ... \rightarrow \hat{\boldsymbol{x}}_{t_n}$
  solve the PF-ODE w/ grad w.r.t. t_0 , i.e. $\hat{\boldsymbol{x}}_0 = \boldsymbol{G}_\theta(\hat{\boldsymbol{x}}_{t_n}, t_n)$
  sample new pure noise $\boldsymbol{z}_g$ and random timestep $t \sim \mathcal{U}(0, T)$
  diffuse sample $\hat{\boldsymbol{x}}_0$ with $\boldsymbol{z}_g$ and Eq. (1), i.e. $\boldsymbol{x}_t = \boldsymbol{q}(\boldsymbol{x}_t|\hat{\boldsymbol{x}}_0)$
  solve the PF-ODE of $\boldsymbol{f}_\psi$ w.r.t. $(t - \Delta t)$ to obtain $\boldsymbol{x}^{\text{fake}}_{t-\Delta t}$
  solve the PF-ODE of $\boldsymbol{F}_\phi$ w.r.t. $(t - \Delta t)$ to obtain $\boldsymbol{x}^{\text{real}}_{t-\Delta t}$
  update θ with $(\boldsymbol{x}^{\text{fake}}_{t-\Delta t}, t - \Delta t)$ and Eq. (7)
  $genIter$ += 1

  // *update discriminator* $\boldsymbol{D}_\tau$
  update τ with $(\boldsymbol{x}^{\text{fake}}_{t-\Delta t}, \boldsymbol{x}^{\text{real}}_{t-\Delta t}, t - \Delta t)$ and Eq. (8)
end while

B. 实现细节

在 ?? 中，我们详细说明了判别器的设计以及两个训练阶段之间的差异。对于附加到判别器主干上的所有可训练头部，以进行文本到图像的实验，我们采用固定的二维设计，遵循 SDXL-Lightning [23] ，它由简单的 4 层 × 4 2D 卷积块组成，步幅为 2，具有 32 个组的组归一化 [76] 以及 SiLU 激活 [10, 54] 层。不同之处在于我们将在网络的不同层附加多个头部。无论是 UNet [57] 的输出、DiT [49] 还是 ViT [5] ，我们都统一地将其重塑为 $[\boldsymbol{B}atch, \boldsymbol{C}hannel, \boldsymbol{H}eight, \boldsymbol{W}idth]$ ，然后将其用作判别器头的输入。对于 SDXL [56] ，我们获取每个块（包括下采样、中间和上采样块）的最后一个 ResNet [9] 的输出，总共得到 7 个判别器头。对于 SD3 系列 [6] 模型，我们获取每个 DiT 块的输出，分别为 SD3-Medium 和 SD3.5-Large 获得 24 和 38 个判别器头。对于 SAM [19] 和 DINOv2 [48] ，我们获取层 3、6、9 和 12 的输出，得到 4 个判别器头。

我们的文本到视频潜在扩散模型的 3D 判别器头由简单的块组成，其中包含 3 个 × 3 个 × 3D 卷积，步幅为 1，3 个 × 2D 卷积，步幅为 2，以及具有 32 组的组归一化和 SiLU 激活层。这与 2D 判别器头的设计相似，只是我们额外插入了几个 3D 卷积层以提取时间相关特征。视频 DiT 主干中特定块的输出被重塑为 $[\boldsymbol{B}atch, \boldsymbol{C}hannel, \boldsymbol{T}ime, \boldsymbol{H}eight, \boldsymbol{W}idth]$ ，并输入到相应的判别器头。在实践中，由于 3D 卷积的计算量，我们每 3 个 DiT 块提取一次特征，从而分别为 2B 和 5B 模型获得 10 个和 14 个判别器头。

在 Tab. 7 中，我们展示了与 DMD2 相比，我们提出方法的训练配置和 GPU 消耗。该表格表明，实际上我们在 GPU 时间较少的情况下，实现了比 DMD2 更好的性能，并且没有对 GPU 内存施加过多的需求。尽管在训练过程中维护更多的网络，我们的实现通过稍后详述的若干优化，达到了可控的内存占用。

	Training Iteration	GPU Number	Elapsed Time	GPU Hours	Micro BatchSize	Max Memory
DMD2	20K	64	60 hours	3840	2	-
DMDX	8K+8K	32	70 hours	2240	4	39.6 GiB
- ADP	8K	32	55 hours	1760	4	39.6 GiB
- ADM	8K	32	15 hours	480	4	24.1 GiB

Table 7. 在 A100 GPU 上与 DMD2 的效率比较。ADP 的耗时已经包含了 ODE 对的收集。

为了减少 GPU 内存占用并提高效率，我们在实现中使用了几种加速技术，包括 完全分片数据并行 (FSDP) [53] 、梯度检查点 [4] 和 BF16 混合精度 [72] 。对于文本到视频的模型，我们进一步整合了 上下文并行 (CP) [83] 和 序列并行 (SP) [15] ，按照 MovieGen [46] 中的常规做法来加速训练和推理。更重要的是，内置于 Pytorch FSDP 中的 CPU 卸载技术对于训练多个网络以节省内存是至关重要的。

启用 CPU 卸载后，每个参数连同相应的梯度和优化器状态可以从 GPU 卸载到 CPU 内存中。结合梯度检查点，正向和反向过程中的 GPU 内存占用几乎与只有一个网络时相同，因为峰值内存现在由每个块的最

大激活决定。这需要增加 CPU 内存以及每次迭代更长的时间。虽然 CPU 内存通常足够且便宜，但我们更有效的方法需要更少的迭代即可达到收敛和满意的结果，并且如 Tab. 7 所示，我们的 DMDX 在一步 SDXL 蒸馏中所用时间比 DMD2 更少。

对于优化器的所有模型（包括文本到图像和文本到视频实验中的生成器、伪造模型和判别器），我们使用不带权重衰减的 AdamW [29] 优化器，beta 参数设为 (0.0, 0.99)，以更及时地捕捉分布的变化。判别器和伪造模型在我们所有实验中的学习率分别固定为 5e-6 和 1e-6。

对于 SDXL，在 ADP 和 ADM 训练过程中生成器的学习率分别为 1e-6 和 1e-7。至于多步 ADM 蒸馏，SD3-Medium LoRA 训练和 SD3.5-Large 完全微调的生成器学习率分别设为 1e-6 和 1e-8。在文本到视频扩散蒸馏的情况下，我们为不同的 8 步 CogVideoX 生成器设置相同的学习率 1e-7。

在所有 ADM 实验中，分类器无指导（CFG）对于真实模型如 DMD 一样是必要的 [85] 。对于 SDXL、SD3-Medium、SD3.5-Large 和 CogVideoX，CFG 值的均匀随机采样范围分别设置为 [6.0, 8.0]、[6.0, 8.0]、[3.0, 4.0] 和 [5.0, 7.0]。选择的范围基于原始基准推理中推荐的 CFG 值，并有所允许的变化。我们观察到此设置足以实现满意的蒸馏性能，而无需广泛调优。

伪模型训练不结合 CFG，使用与扩散模型标准预训练相同的损失函数，只是我们没有设置任何 dropout。对于噪声参数化模型，预测目标是噪声，而对于速度参数化模型，预测目标是速度。

C. 主要结果

C.1. 高效图像合成

Fig. 6 定性地将我们的方法与其他最先进的蒸馏技术在 SD3 [6] 系列模型上进行了比较。结果表明，我们的方法在色彩、细节、结构和图文对齐方面与原始模型具有竞争力，同时优于包括 TSCD、PCM [69] 、Flash [3] 和 LADD [60] 在内的其他方法。

C.2. 高效视频合成

Tabs. 8 and 9 介绍了 VBench [14] 在基础模型和 CogVideoX [83] 的少量步生成器上的结果的详细信息。在 Figs. 10 to 15 中，我们展示了若干案例，用于比较我们的 CogVideoX [83] 生成器和基准模型。结果表明，即使在某些情况下存在语义增强，例如 Fig. 10 中的光线变化和 Fig. 13 中羊的运动，我们的 8 步生成器在语义上通常与原始模型相当。而在成像质量方面，generators with CFG are generally more detailed and have more delicate textures 比那些没有 CFG 的要好。细节上的不足体现在例如 Fig. 14 中的手略显粗糙以及手指数目不正确，而有 CFG 的则更自然。此外，generator without CFG is also much higher in color contrast 视觉上有时显得过于鲜艳以致不够真实。这些证明了 CFG 对文本到视频模型的重要性，这可能无法通过定量指标完全体现。

至于 Fig. 7 中展示的对抗蒸馏的消融，其他基线设置的两个主要问题是结构和模糊性。当像在 Rectified Flow [27] 中使用 MSE 损失进行单次重流过程时，显然它在生成结构上可见的图像时遇到了困难。而将 SAM [19] 模型切换为 DINOv2 [48] 时，我们可以清晰地看到图中机器人和面部的结构崩溃，这在意料之外，可能是因为其输入分辨率仅为 518px，而我们生成的图像均为 1024px，需要在输入前调整大小。另一种可能的解释是，SAM 用于实例分割的先验知识比 DINOv2 用于判别自监督学习所提供的丰富，促进了局部细粒度细节的生成。在增加像素空间 λ_2 的权重时所遇到的结构问题类似，而减少其权重则会导致在图中明显可见的显著模糊，因此我们建议设置 $\lambda_1 = 0.85, \lambda_2 = 0.15$ 是一个合理的配置。

在 Fig. 8 中，我们为分数蒸馏的消融研究提供了定性比较。与仅使用 ADP 的基线相比，我们可以看到 ADM 蒸馏确实作为微调过程，在颜色、细节和最显著结构方面优化了生成器。虽然独立的 ADM 也能产生有效的生成器，但正如 [23, 85] 所观察到的那样，在单步生成中仍然存在噪声伪影，而通过我们的 ADP 可以很好地解决这个问题。值得注意的是，可视化结果表明，不集成 ADP 使用 DMD 损失会引入相当严重的噪声伪影。与单独使用 ADM 相比，其定性劣势比定量结果中观察到的差距更加明显。使用 ADP 后，DMD 损失产生相对较好的结果，但在视觉保真度和结构完整性方面仍不如 ADM。这表明其分布匹配能力比 ADM 弱，这与我们在 Sec. 5.3 中的定量结果分析一致。

此外，我们展示了在 Fig. 9 中与 DMD2 相比额外随机整理的多种种子样本，清楚地表明我们的图像在纹理、颜色、亮度、对比度和结构构成方面表现出更丰富的变化。

考虑到许多当前方法利用基础模型生成的数据作为辅助 [44] ，我们的扩散模型加速方法可以大大加快这一过程，从而有利于许多下游任务，如识别 [77] 、检测 [42] 、检索 [31, 41] 、领域适应 [32, 62] 等。或者，我们可以训练 LoRA 以获得加速插件，增强图像 [33] 或视频 [80] 生成的定制垂直模型的效率。

D. 提示列表

以下我们列出了本文所示生成内容所使用的文本提示（从上到下，从左到右）。注意，由于像 SDXL-Base [56] 这样的模型仅使用 CLIP [52] 作为文本编码器，它仅支持最多 77 个标记，对于一些较长的提示，响应和文本-图像对齐可能会不够充分，并且在理解方面也有其有限的能力。

We use the following prompts for Fig. 5:

- 一个美丽的女人面向镜头，自信地微笑，彩色长发，钻石项链，深红色的嘴唇，中景，非常细致，逼真，杰作。
- 一只猫头鹰安静地栖息在古老森林深处的一根扭曲的树枝上。它锐利的黄色眼睛敏锐而警觉。

Method	Step	NFE	Final Score	Quality Score	Semantic Score	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
ADM	8	8	78.58	80.82	69.62	96.72	96.55	97.01	98.14	48.61	57.80	65.28
+Longer Training × 2	8	8	80.76	83.03	71.69	96.58	96.71	98.12	97.68	73.33	57.90	65.72
ADM w/ CFG	8	16	79.86	80.93	75.56	96.16	96.96	96.86	97.69	54.44	59.78	63.18
+Longer Training × 2	8	16	81.79	83.00	76.94	96.83	96.90	98.51	98.07	63.05	61.03	64.62
CogVideoX-2b	100	200	80.03	80.80	76.97	92.53	95.22	97.79	97.00	69.44	60.38	60.69
ADM	8	8	82.06	83.22	77.42	96.42	96.87	96.96	97.69	68.88	61.17	69.01
ADM w/ CFG	8	16	80.98	82.16	76.25	96.15	96.59	95.99	98.57	56.66	61.01	68.68
CogVideoX-5b	100	200	81.22	81.78	78.98	92.52	96.68	98.34	96.97	70.55	61.67	61.88

Table 8. VBench [14] 的详细结果包括总体得分和每个质量维度的单独得分。

Method	Step	NFE	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
ADM	8	8	83.97	47.19	87.40	77.79	62.93	42.64	24.16	22.35	25.27
+Longer Training × 2	8	8	87.84	56.53	85.00	80.28	69.52	44.33	23.15	22.60	25.11
ADM w/ CFG	8	16	89.55	64.78	92.60	82.31	62.61	52.73	24.31	24.46	26.12
+Longer Training × 2	8	16	91.67	71.58	92.20	82.01	71.79	50.26	23.54	24.54	26.30
CogVideoX-2b	100	200	80.01	67.23	98.60	89.98	49.05	68.60	24.04	25.37	25.68
ADM	8	8	92.94	65.89	95.80	84.97	72.92	56.06	22.63	23.64	26.17
ADM w/ CFG	8	16	89.41	69.89	97.00	71.35	81.26	53.90	21.48	23.79	25.92
CogVideoX-5b	100	200	87.64	67.34	99.60	83.93	68.24	56.35	25.16	25.82	27.79

Table 9. VBench [14] 关于每个语义维度的单独评分的详细结果。

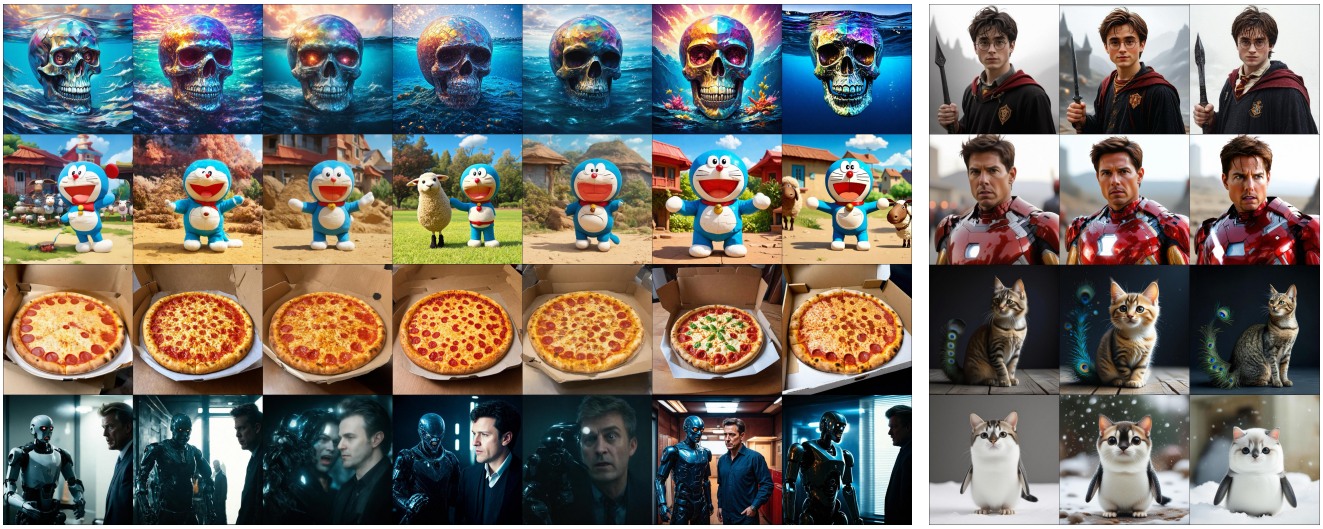

Figure 6. 关于 LoRA 微调 SD3-Medium 和完全微调 SD3.5-Large 的定性结果。

- 一只年轻的獾小心翼翼地嗅着一朵黄玫瑰，背景中潜伏着一只狮子。
- 一辆皮卡车正在爬山道转弯处。

We use the following prompts for Fig. 6:

- 一张巨型钻石骷髅在海洋中的照片，呈现出鲜艳的色彩和细致的纹理。
- 一张来自“小羊肖恩”的哆啦 A 梦静态画面，由阿德曼动画公司制作。
- 一个披萨在披萨盒内展示。
- 电影剧照：一个男人和一个机器人处于惊恐的时刻，电影剧照，电影构图，电影灯光，由埃德加 · 赖特和大卫 · 林奇制作
- 哈利 · 波特作为《上古卷轴 V：天际》中的角色
- 电影剧照中由汤姆 · 克鲁斯饰演的《复仇者联盟》中的钢铁侠
- 一张获奖的精美图片，展示了一只在暗色背景前的可爱猫咪。这只猫是猫-孔雀杂交种，拥有孔雀尾巴和身上有短的孔雀羽毛。绒毛蓬松，极其细致，令

Figure 7. 对抗蒸馏消融研究的定性比较。

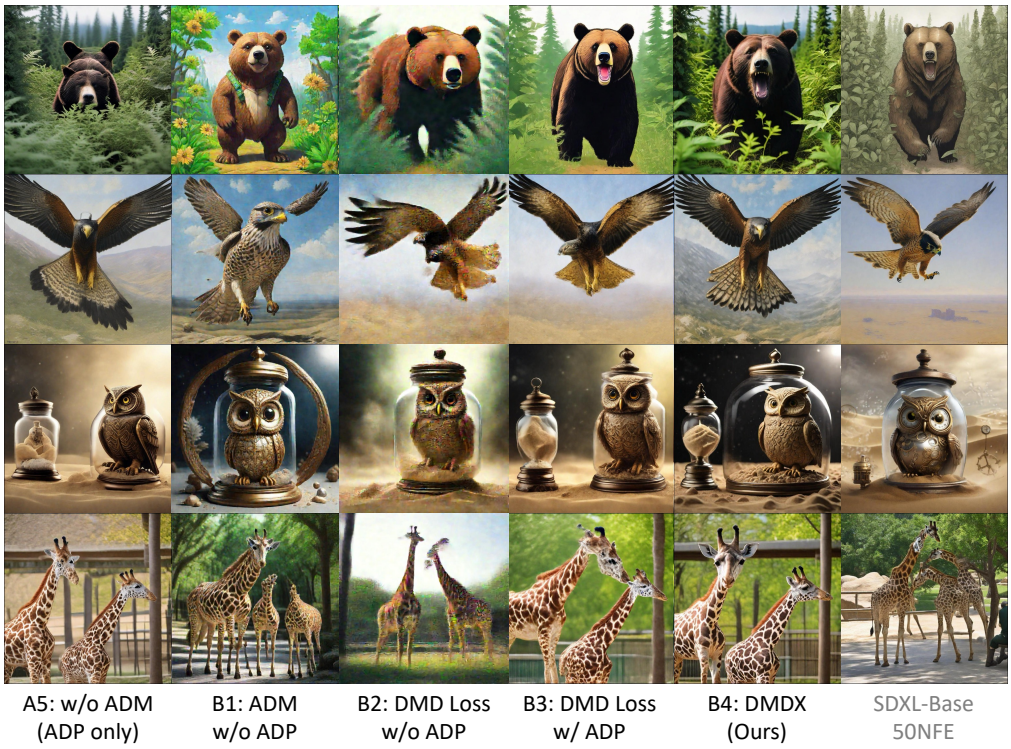

Figure 8. 关于得分蒸馏消融研究的定性比较。

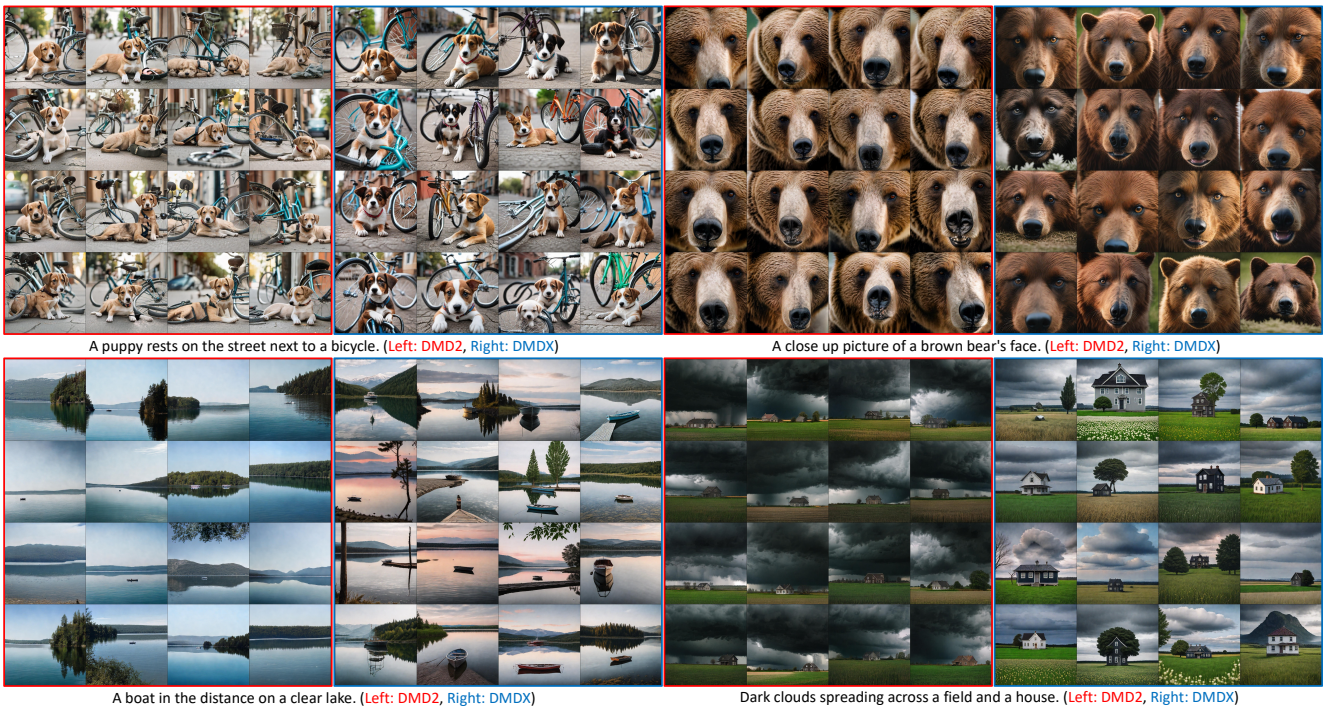

Figure 9. 使用 DMD2 进行定性多样性比较。

人惊叹，高质量，氛围灯光。

- 一种可爱的动物，是企鹅和猫的混合体

We use the following prompts for Fig. 7:

- 一个色彩斑斓的锡制玩具机器人在瑞士阿尔卑斯山美丽的花卉草坪附近的小路上运行蒸汽发动机，背景是山景全景图，整个画面以长镜头拍摄，带有运动模糊和景深效果。
- Leighann Vail 的肖像画。
- 一张由 Stanley Lau 和 Artgerm 创作的，Stefan Kostic 科幻风格的，带有水晶翅膀的机械天使女性的照片。
- 一幅描绘印度夏季小径的画作，背景是壮丽的傍晚天空，落日余晖和低沉的雷云。

We use the following prompts for Fig. 8:

- 一只熊嘴里叼着一株植物穿过一片灌木丛。
- 伊利亚 · 列宾、菲尔 · 黑尔和肯特 · 威廉姆斯用高度细致的画作描绘了一只飞翔的猎鹰。
- 一个蒸汽朋克怀表猫头鹰被困在埋在沙中的玻璃罐中，周围环绕着沙漏和旋转的薄雾。
- 一些长颈鹿正在动物园的展览区四处走动。

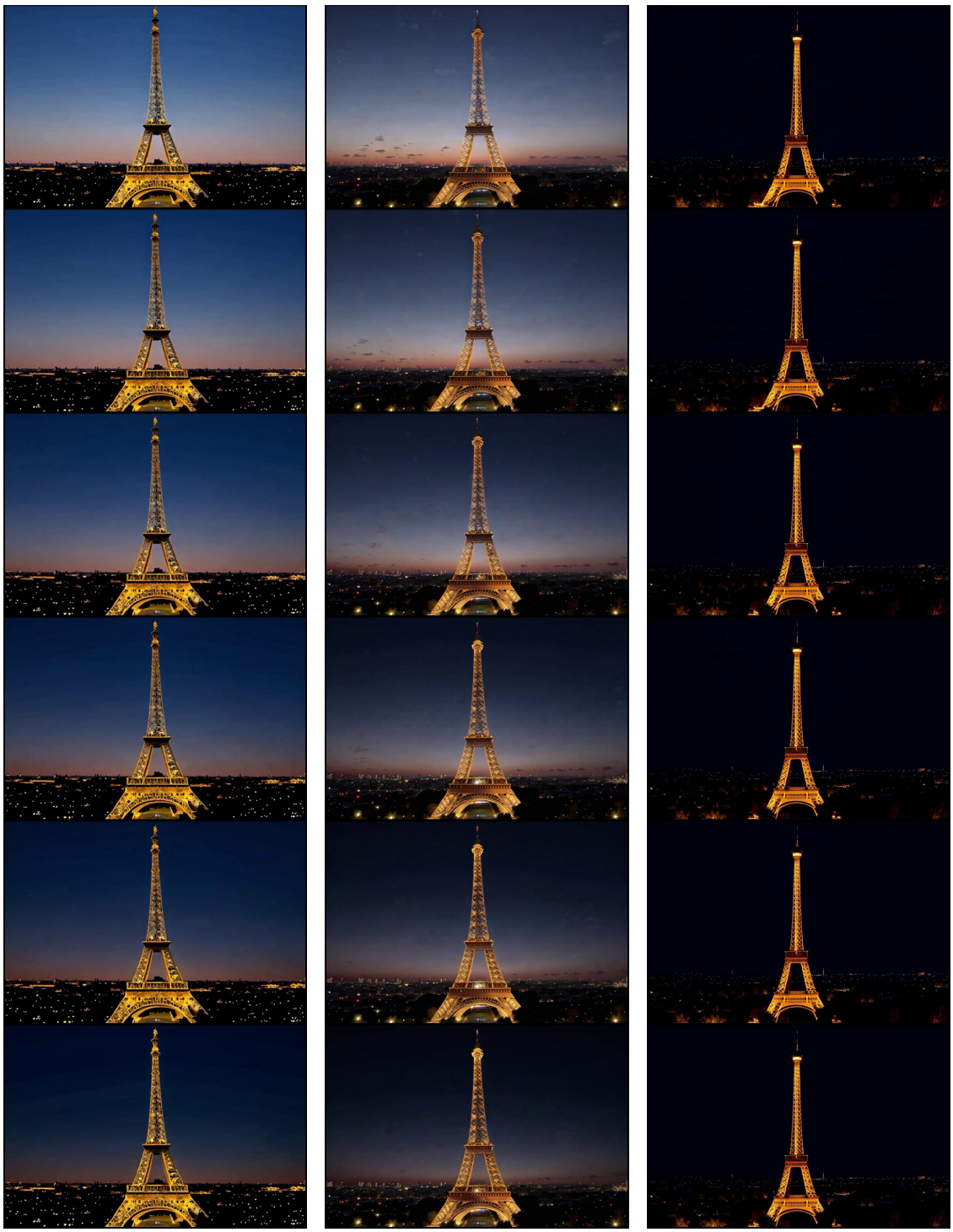

Figure 10. CogVideoX-2b 生成器的定性比较。随机种子已固定。提示：一个延时序列捕捉到了标志性埃菲尔铁塔从白天到晚上的变换。铁塔高高耸立，以其原始的金色光芒呈现出宏伟之姿，gradually transitions into a silhouette against the twilight sky 。当太阳落下，城市灯光开始闪烁，整个巴黎景色被温暖的光芒笼罩。塔的复杂铁制框架结构变得更加清晰，其阴影在马尔斯广场上拉长。背景中包括塞纳河和巴黎的屋顶，为场景增添了深度和背景。随着夜幕降临，埃菲尔铁塔被自身的灯光照亮，化作巴黎的一座灯塔，在繁星点缀的背景下闪烁。

Figure 11. CogVideoX-2b 生成器的定性比较。随机种子已固定。提示：一棵充满生机的橡树，装饰着节日万圣节的装饰品，矗立在郊区的后院中。粗壮坚固的树干支撑着各种装饰品。树枝上挂着明亮的橙色和黑色气球、诡异的蜘蛛网以及飘扬的幽灵。一个 large, carved pumpkin sits at the base ，复杂的面容在温暖、宜人的光芒中闪烁。这个场景以整齐修剪的树篱和一条通向一栋别致房子的道路为背景，沐浴在柔和的秋日阳光中。

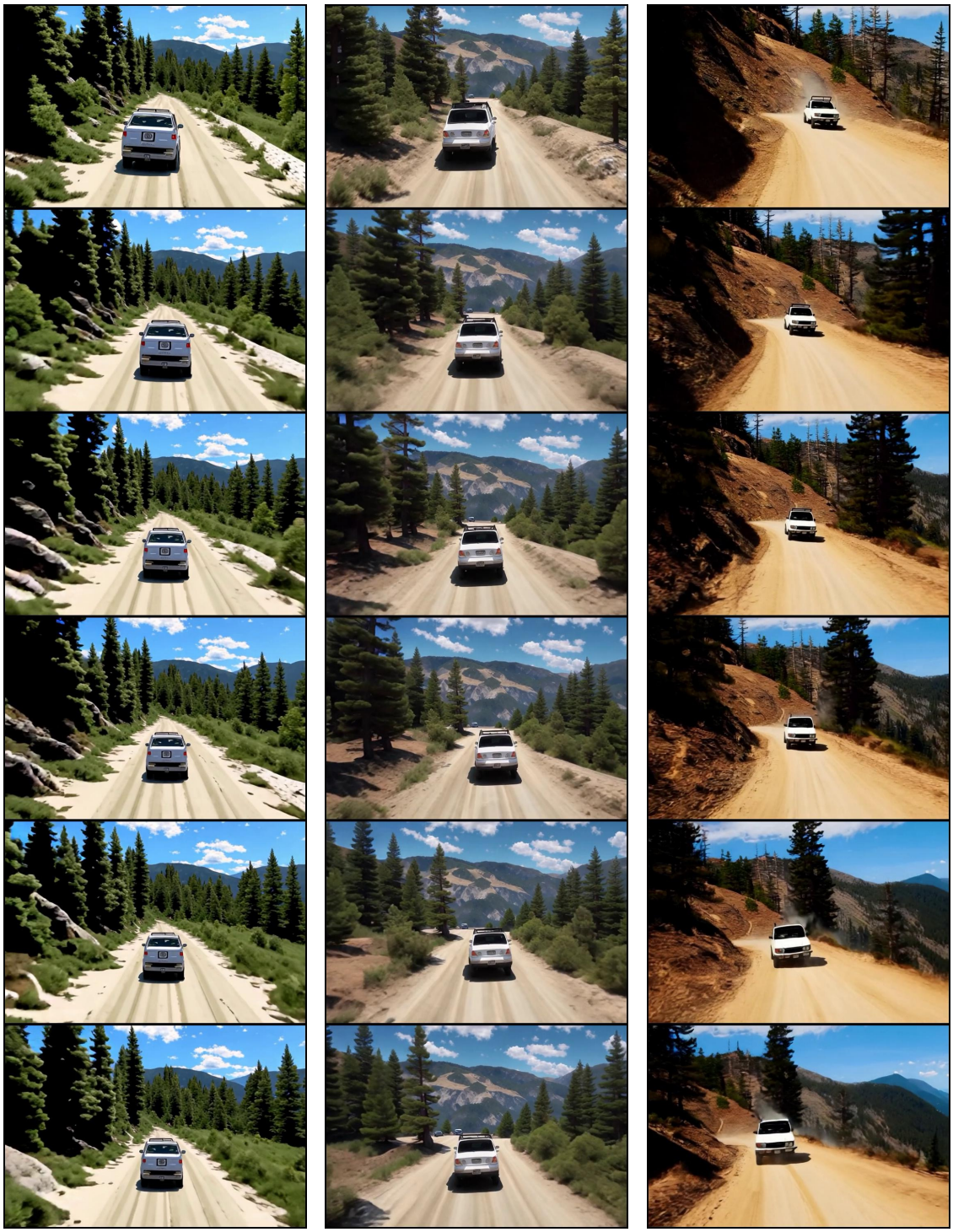

Figure 12. 关于 CogVideoX-2b 生成器的定性比较。随机种子已固定。提示词：一辆带黑色行李架的 camera follows behind a white vintage SUV 在陡峭的山坡上加速行驶，这条陡峭的土路被松树环绕，车轮扬起尘土，阳光洒在正在沿土路疾驰的 SUV 上，为这个场景增添了一抹暖色。土路在远方轻轻地弯曲，视野中没有其他车辆。路两边的树木是红杉，散落着绿意。汽车从后面被看到，轻松地沿着弯道行驶，仿佛在崎岖的地形上越野行驶。土路本身被陡峭的山丘和山脉包围，天空晴朗，白云缭绕。

Figure 13. CogVideoX-5b 生成器的定性比较。随机种子已固定。提示：一只毛茸茸的白羊站在郁郁葱葱的绿色草地上，它的羊毛在温暖的午后阳光下闪闪发光。场景转换到 a close-up of the sheep's gentle face ，它的大而好奇的眼睛和柔软、微微颤动的耳朵吸引了注意。背景是点缀着野花的连绵起伏的山丘和晴朗的蓝天。羊随后平静地吃草，它的动作缓慢而审慎，微风轻轻拂过草地。最后，the sheep looks up, framed by the picturesque landscape ，体现了宁静和自然的简单之美。

Figure 14. CogVideoX-5b 生成器的定性比较。随机种子已固定。提示: 格温 · 史黛西把标志性的金色头发绑成马尾，坐在一个舒适的阳光明媚的房间里，专心致志地读着一本厚重的皮革装订的书。她穿着一套休闲但时尚的服装：淡蓝色毛衣、深色牛仔裤和黑色踝靴。镜头从她轻轻翻页的手开始，露出她精心修饰的指甲。当 the camera tilts up, it captures her focused expression ，她的眼睛好奇而专注地扫过文字。温暖的阳光透过附近的窗户照射进来，给她的脸庞投射出柔和的光芒，突显出她宁静而专心的神情。场景 ends with a close-up of her thoughtful smile ，暗示了一个发现或反思的时刻。

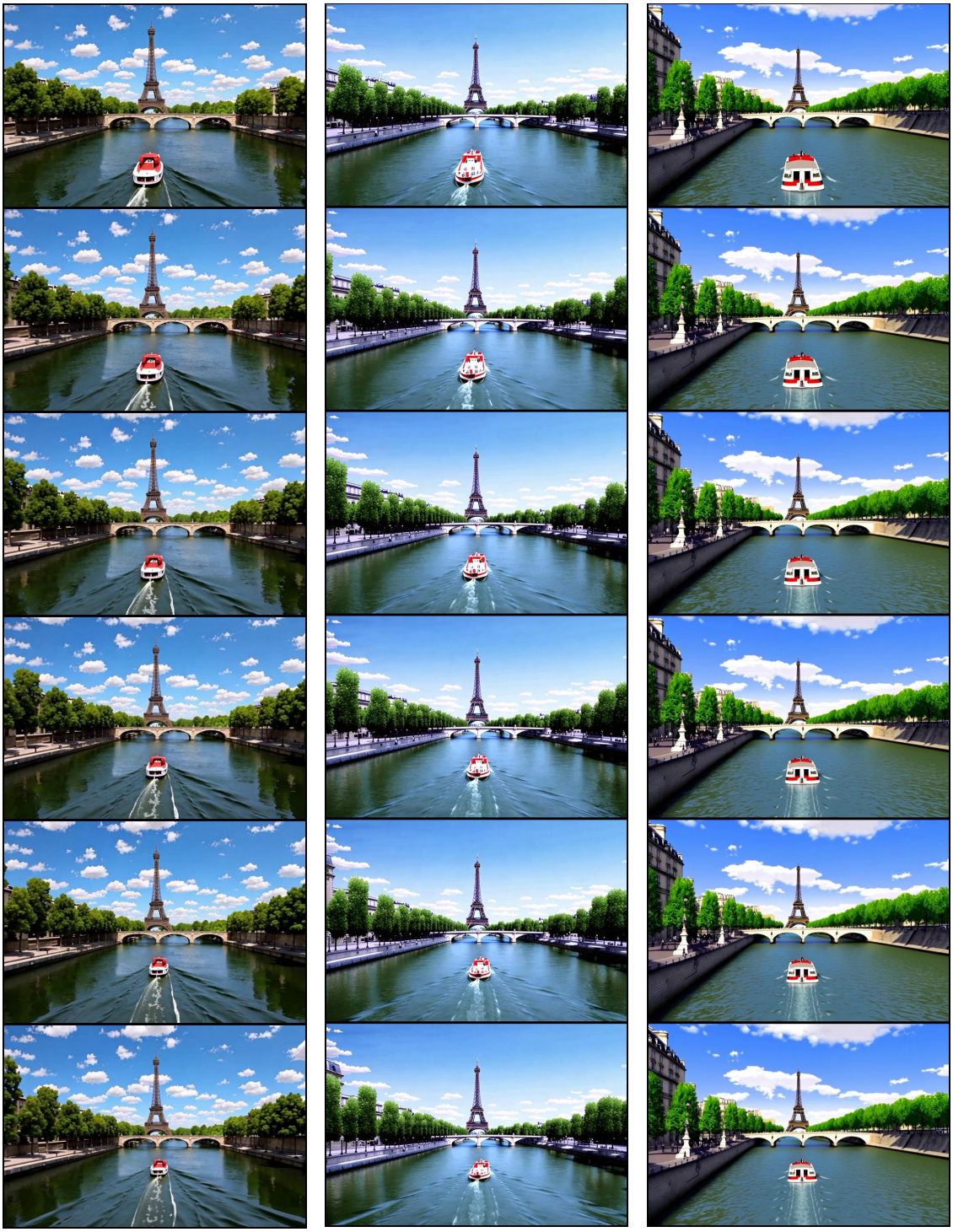

Figure 15. CogVideoX-5b 生成器的定性比较。随机种子已固定。提示词：一艘迷人的小船，红白色的船体，悠闲地航行在宁静的塞纳河上，船尾扬起的轻柔波纹在水中荡漾。标志性的埃菲尔铁塔雄伟地矗立在背景中，蓝天朗朗，白云朵朵衬托其美姿。当 the camera zooms out, the scene expands to reveal lush green trees lining the riverbanks ，古朴的巴黎建筑，经典的建筑风格，以及在鹅卵石小径上漫步的行人都一一显现。小船继续它宁静的旅程，经过装饰精美的灯柱点缀的优雅石桥，捕捉到了巴黎一个平和日子的精髓。