

# 液晶模组-L氧RA：澳大利亚通用S桌子-D扩散A加速中号欧都莱

罗四面<sup>\*,1</sup> 谭益勤<sup>\*,1</sup> 苏拉杰·帕蒂尔<sup>†,2</sup> 顾丹尼尔<sup>†</sup> 帕特里克·冯·普拉顿<sup>2</sup>  
 阿波利纳里奥·帕索斯<sup>2</sup> 黄龙波<sup>1</sup> 李健<sup>1</sup> 赵航<sup>1</sup>  
<sup>1</sup>清华大学信息科学研究院 <sup>2</sup>抱脸  
 {luosm22、tyq22}@mails.tsinghua.edu.cn  
 {帕特里克·苏拉吉·阿波利纳里奥}@huggingface.co {  
 dgu8957}@gmail.com  
 {龙博皇、lijian83、hangzhao}@tsinghua.edu.cn

## ABSTRACT

潜在一致性模型 (LCM) (Luo 等人, 2023) 在加速文本到图像生成任务、以最少的推理步骤生成高质量图像方面取得了令人印象深刻的性能。LCM 是从预训练的潜在扩散模型 (LDM) 中提取出来的, 只需要 ~32 个 A100 GPU 训练小时。该报告进一步扩展了 LCM 在两个方面的潜力: 首先, 通过将 LoRA 蒸馏应用于稳定扩散模型, 包括 SD-V1.5 (Rombach et al., 2022)、SSD-1B (Segmind., 2023) 和 SDXL (Podell 等人, 2023), 我们将 LCM 的范围扩展到更大的模型, 内存消耗显著减少, 实现了卓越的图像生成质量。其次, 我们将通过 LCM 蒸馏获得的 LoRA 参数确定为通用稳定扩散加速模块, 命名为 LCM-LoRA。LCM-LoRA 可直接插入各种稳定扩散微调模型或 LoRA 未经训练, 从而代表了适用于各种图像生成任务的普遍适用的加速器。与之前的数值 PF-ODE 求解器如 DDIM (Song et al., 2020)、DPM-Solver (Lu et al., 2022a;b) 相比, LCM-LoRA 可以被视为插件式神经 PF-ODE 求解器具有很强的泛化能力。项目页面: <https://github.com/luosiallen/>

潜在一致性模型。

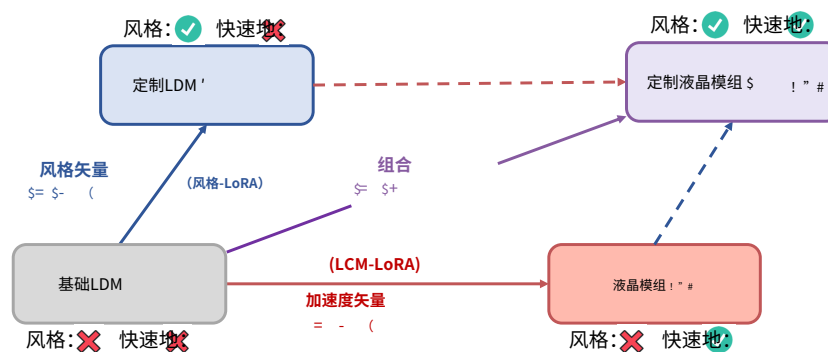


图 1: LCM-LoRA 概述。通过将 LoRA 引入 LCM 的蒸馏过程, 我们显著减少了蒸馏的内存开销, 这使得我们能够在有限的资源下训练更大的模型, 例如 SDXL 和 SSD-1B。更重要的是, 通过 LCM-LoRA 训练得到的 LoRA 参数

(“加速向量”) 可以直接与通过对特定样式数据集进行微调获得的其他 LoRA 参数 (“样式向量”) 相结合。无需任何训练, 通过加速度向量和风格向量的线性组合获得的模型就能够以最少的采样步骤生成特定绘画风格的图像。

主要作者  
 † 核心贡献者

## 1 我简介

潜在扩散模型 (LDM) (Rombach 等人, 2022) 对于从文本和草图等各种输入生成高度详细且富有创意的图像至关重要。尽管取得了成功, 但 LDM 固有的缓慢反向采样过程阻碍了实时应用程序, 损害了用户体验。当前的开源模型和加速技术尚未弥合标准消费级 GPU 上实时生成的差距。加速 LDM 的努力通常分为两类: 第一类涉及先进的 ODE 求解器, 如 DDIM (Song et al., 2020)、DPM-Solver (Lu et al., 2022a) 和 DPM-Solver++ (Lu et al., 2022a)。2022b), 以加快生成过程。第二个策略涉及对 LDM 进行提炼, 以简化其运作。ODE-Solver 方法尽管减少了所需的推理步骤数量, 但仍然需要大量的计算开销, 特别是在合并无分类器指导时 (Ho & Salimans, 2022)。与此同时, Guided-Distill (Meng et al., 2023) 等蒸馏方法虽然很有前景, 但由于其密集的计算要求而面临实际限制。在 LDM 生成的图像中寻求速度和质量之间的平衡仍然是该领域的一个挑战。

最近, 受一致性模型 (CM) (Song 等人, 2023) 的启发, 潜在一致性模型 (LCM) (Luo 等人, 2023) 出现, 作为图像生成中采样缓慢问题的解决方案。LCM 通过将反向扩散过程视为增强概率流 ODE (PF-ODE) 问题来处理它。他们创新地预测了潜在空间中的解决方案, 绕过了通过数值常微分方程求解器迭代解决方案的需要。这使得高分辨率图像的合成非常高效, 仅需要 1 到 4 个推理步骤。此外, LCM 在蒸馏效率方面也很突出, 只需 32 个 A100 训练小时即可进行最少步骤的推理。

在此基础上, 潜在一致性微调 (LCF) (Luo et al., 2023) 被开发为一种无需从教师扩散模型开始即可微调预训练 LCM 的方法。对于专门的数据集 (例如动画、逼真或奇幻图像的数据集), 需要额外的步骤, 例如采用潜在一致性蒸馏 (LCD) (Luo 等人, 2023) 将预先训练的 LDM 蒸馏为 LCM 或使用 LCF 直接微调 LCM。然而, 这种额外的训练可能会成为跨不同数据集快速部署 LCM 的障碍, 从而提出了是否可以对自定义数据集进行快速、免训练的推理的关键问题。

为了回答上面的问题, 我们引入 LCM-LoRA, A 通用免训练加速模块可以直接插入各种稳定扩散 (SD) (Rombach 等人, 2022) 微调模型或 SD LoRA (Hu 等人, 2021) 中, 以支持以最少的步骤进行快速推理。与早期的数值概率流 ODE (PF-ODE) 求解器相比, 例如 DDIM (Song et al., 2020)、DPM-Solver (Lu et al., 2022a) 和 DPM-Solver++ (Lu et al., 2022b), LCM-LoRA 代表了一类新型的基于神经网络的 PF-ODE 求解器模块。它展示了跨各种微调 SD 模型和 LoRA 的强大泛化能力。

## 2 R兴高采烈瓦奥克

一致性模型宋等人。(2023) 展示了一致性模型 (CM) 的巨大潜力, 这是一类新型的生成模型, 可以在不牺牲输出质量的情况下提高采样效率。这些模型采用一致性映射技术, 可以巧妙地将常微分方程 (ODE) 轨迹上的点映射到其原点, 从而实现快速一步生成。他们的研究专门针对 ImageNet 64x64 (Deng 等人, 2009) 和 LSUN 256x256 (Yu 等人, 2015) 上的图像生成任务, 证明了 CM 在这些领域的有效性。进一步推进该领域, 罗等人。(2023) 开创了潜在一致性模型 (液晶模组) 在文本到图像合成领域。通过将引导反向扩散过程视为增强概率流 ODE (PF-ODE) 的分辨率, LCM 可以熟练地预测此类 ODE 在潜在空间中的解。这种创新方法显着减少了迭代步骤的需求, 从而能够从文本输入快速生成高保真图像, 并为 LAION-5B-Aesthetics 数据集的最先进性能设定新标准 (Schuhmann 等人, 2017)。, 2022)。

参数高效微调参数高效微调 (PEFT) (Houlsby 等人, 2019) 可以为特定任务定制预先存在的模型, 同时限制模型的数量

需要重新训练的参数。这减少了计算负载和存储需求。在 PEFT 保护下的各种技术中，低秩适应 (LoRA) (Hu 等人, 2021) 脱颖而出。LoRA 的策略包括通过低秩矩阵的集成来训练最小的参数集，这些参数简洁地表示了模型权重进行微调所需的调整。实际上，这意味着在特定于任务的优化期间，仅学习这些矩阵，并且大部分预训练的权重保持不变。因此，LoRA 显着减少了需要修改的参数数量，从而提高了计算效率，并允许使用相当少的数据来细化模型。

预训练模型中的任务算法任务算术 (Ilharco et al., 2022; Ortiz-Jimenez et al., 2023; Zhang et al., 2023) 已成为增强预训练模型能力的一种著名方法，为以下任务提供了一种经济有效且可扩展的策略：直接编辑权重空间。通过将不同任务的微调权重应用于模型，研究人员可以提高模型在这些任务上的表现，或者通过否定它们来诱发遗忘。尽管有其前景，但对任务算术的全部潜力及其背后原理的理解仍然是积极探索的领域。

### 3 LCM-L氧RA

#### 3.1 升氧RAD蒸馏用于液晶模组

潜在一致性模型 (LCM) (Luo 等人, 2023) 使用单阶段引导蒸馏方法进行训练，利用预先训练的自动编码器的潜在空间将引导扩散模型蒸馏为 LCM。此过程涉及求解增强概率流 ODE (PF-ODE)，这是一种数学公式，可确保生成的样本遵循产生高质量图像的轨迹。蒸馏的重点是保持这些轨迹的保真度，同时显着减少所需的采样步骤数量。该方法包括诸如跳跃步骤技术之类的创新，以加速收敛。算法1提供了LCD的伪代码。

---

#### 算法1潜在稠度蒸馏 (LCD) (Luo 等人, 2023)

---

输入：数据集  $D$ ，初始模型参数  $\theta$ ，学习率  $\eta$ ，常微分方程求解器  $\Psi(\cdot, \cdot, \cdot, \cdot, \cdot)$ ，距离度量  $d(\cdot, \cdot)$ ，EMA 率  $\mu$ ，噪声表  $\alpha(t), \sigma(t)$ ，指导尺度  $[W_{\text{分钟}}, W_{\text{最大限度}}]$ ，跳过间隔  $k$ ，和编码器  $\mathcal{Z}(\cdot)$  将训练数据编码到潜在空间中： $D = \{(z, c) | z = \mathcal{Z}(X), (X, c) \in D\}$   $\theta \leftarrow \theta$

重复

  样本  $(z, c) \sim D, n \sim U[1, N-k]$  和  $\omega \sim [\omega_{\text{分钟}}, \omega_{\text{最大限度}}]$  样本  
   $z_{t_{n+k}} \sim N(\alpha(t_{n+k}), \sigma(t_{n+k}))$  我  
   $\hat{z}_{\Psi, \omega} \leftarrow z_{t_{n+k}} + (1+\omega)\Psi(z_{t_{n+k}}, t_{n+k}, t_n, C) - \omega\Psi(z_{t_{n+k}}, t_{n+k}, t_n, \emptyset)$   $L(\theta, \theta; \Psi) \leftarrow d(\mathcal{F}_{\theta}(z_{t_{n+k}}, C, t_{n+k}), \mathcal{F}_{\hat{\theta}}(\hat{z}_{\Psi, \omega}, C, t))$   $t_n \quad n$   
   $\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta, \theta)$   
   $\theta \leftarrow \text{停止梯度}(\mu\theta + (1-\mu)\theta)$  直

到收敛

---

由于潜在一致性模型 (LCM) 的蒸馏过程是在预先训练的扩散模型的参数之上进行的，因此我们可以将潜在一致性蒸馏视为扩散模型的微调过程。这使我们能够采用参数高效的微调方法，例如 LoRA (低秩适应) (Hu 等人, 2021)。LoRA 通过应用低秩分解来更新预训练的权重矩阵。给定一个权重矩阵  $\bar{W}_0 \in \mathbb{R}^{d \times k}$ ，更新表示为  $\bar{W}_0 + \Delta \bar{W} = \bar{W}_0 + \text{学士}$ ，在哪里  $\mathcal{B} \in \mathbb{R}^{d \times r}$ ,  $A \in \mathbb{R}^{r \times k}$ ，以及排名  $r \leq \min(d, k)$ 。在训练中， $\bar{W}_0$  保持不变，并且梯度更新仅应用于  $A$  和  $\mathcal{B}$ 。输入的修改前向传播  $X$  是：

$$H = \bar{W}_0 X + \Delta \text{维} x = \bar{W}_0 X + \text{巴克斯}。 \quad (1)$$

在这个等式中， $H$  表示输出向量，并且输出  $\bar{W}_0$  和  $\Delta \bar{W} = \text{学士}$  乘以输入后相加  $X$ 。通过将完整参数矩阵分解为两个低秩矩阵的乘积，LoRA 显着减少了可训练参数的数量，从而降低了内存使用量。

表3.1 完整参数总数对比

## 4 步推理

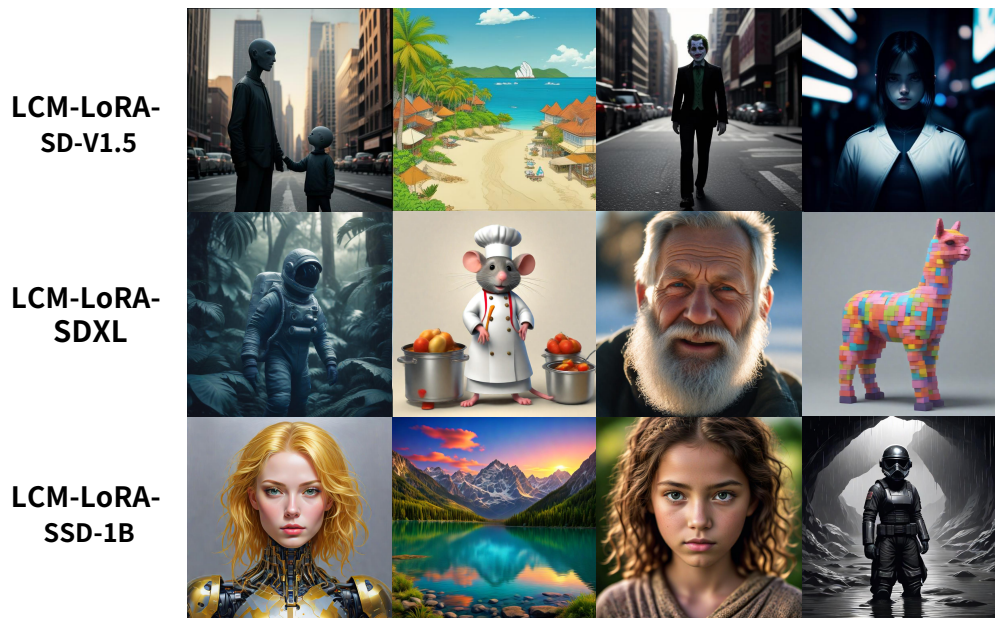


图 2: 使用从不同预训练扩散模型中提取的潜在一致性模型生成的图像。我们生成  $512 \times 512$  分辨率图像使用 LCM-LoRA-SD-V1.5 和 1024 的 512 分辨率图像使用 LCM-LoRA-SDXL 和 LCM-LoRA-SSD-1B 提供 1024 分辨率图像。我们使用固定的无分类器指导量表  $\omega=7.5$  适用于蒸馏过程中的所有模型。所有图像均通过 4 步采样获得。

使用 LoRA 技术时具有可训练参数的模型。显然，通过在 LCM 蒸馏过程中结合 LoRA 技术，可训练参数的数量显著减少，有效降低了训练的内存需求。

模型	SD-V1.5	SSD-1B	SDXL
# 完整参数	0.98B	1.3B	3.5B
# LoRA 可训练参数	67.5M	105M	197M

表 1: SD-V1.5 (Rombach 等人, 2022)、SSD-1B (Segmind., 2023) 和 SDXL (Podell 等人, 2023) 的 LoRA 的完整参数数和可训练参数数。

罗等人。(2023)主要提炼了基础稳定扩散模型，如SD-V1.5和SD-V2.1。我们将此蒸馏过程扩展到更强大的模型，具有增强的文本到图像功能和更大的参数数量，包括 SDXL (Podell 等人, 2023) 和 SSD-1B (Segmind., 2023)。我们的实验表明 LCD 范式可以很好地适应更大的模型。不同模型的生成结果如图2所示。

## 3.2 LCM-LoRA作为通用加速中号欧都莱

基于 LoRA 等参数高效的微调技术，人们可以在大幅降低内存需求的情况下微调预训练模型。在 LoRA 框架内，得到的 LoRA 参数可以无缝集成到原始模型参数中。在第 3.1 节中，我们演示了采用 LoRA 进行潜在一致性模型 (LCM) 蒸馏过程的可行性。另一方面，可以针对特定的面向任务的应用程序对定制数据集进行微调。现在大量的微调参数可供选择和使用。我们发现 LCM-LoRA 参数可以直接与在特定风格的数据集上微调的其他 LoRA 参数相结合。这种合并产生了一个模型，能够以最少的采样步骤生成特定风格的图像，而无需

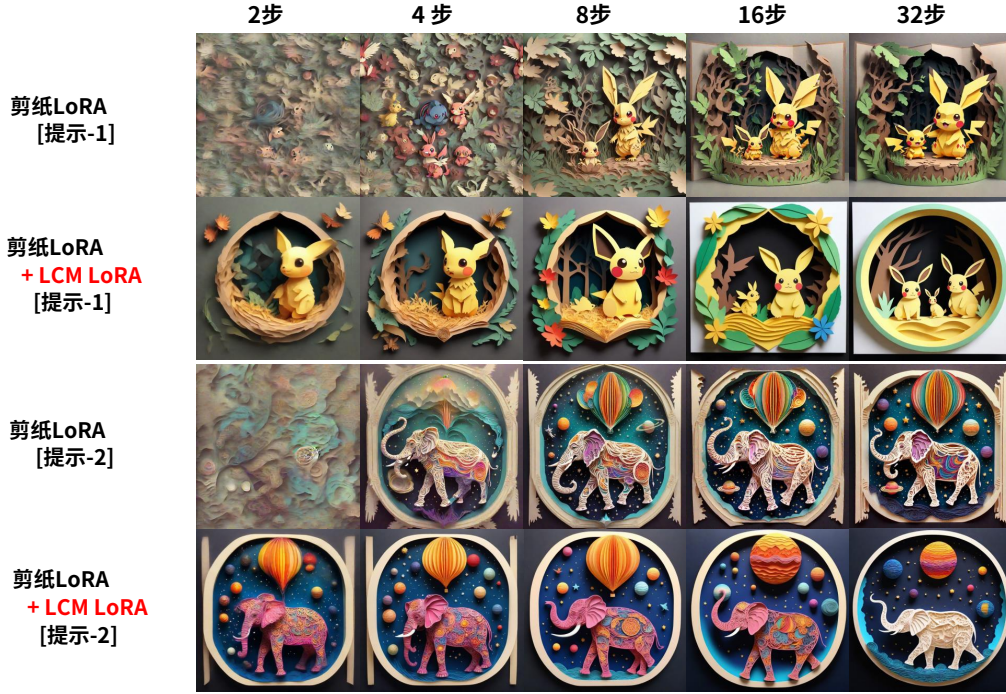


图3：特定风格LoRA参数的生成结果以及与LCM-LoRA参数的组合。我们使用 SDXL 作为基础模型。所有图像均为  $1024 \times 1024$  分辨率。我们选择在特定绘画风格数据集上微调的 LoRA 参数，并将其与 LCM-LoRA 参数相结合。我们比较这些模型在不同采样步骤生成的图像质量。对于原始 LoRA 参数，我们使用 DPM-Solver++ (Lu et al., 2022b) 采样器和无分类器指导量表  $\omega=7.5$ 。对于 LCM-LoRA 与特定风格 LoRA 结合后获得的参数，我们使用 LCM 的多步采样器。我们用  $\lambda_1=0.8$  和  $\lambda_2=1.0$  对于组合。

是否需要任何进一步的培训。如图1所示，将 LCM-LoRA 微调参数表示为  $\tau_{\text{液晶模组}}$ ，被识别为“加速度向量”，LoRA 参数在定制数据集上微调为  $\tau$ ，即“风格向量”，我们发现生成定制图像的 LCM 可以获得为

$$\theta_{\text{液晶模组}} = \theta_{\text{前}} + \tau' \quad (2)$$

在哪里

$$\tau_{\text{液晶模组}} = \lambda_1 \tau + \lambda_2 \tau_{\text{液晶模组}} \quad (3)$$

是加速度矢量的线性组合  $\tau_{\text{液晶模组}}$  和风格向量  $\tau$ 。这里  $\lambda_1$  和  $\lambda_2$  是超参数。特定风格 LoRA 参数的生成结果及其与 LCM-LoRA 参数的组合如图3所示。请注意，我们不对组合参数进行进一步的训练。

#### 4C 结论

我们推出了 LCM-LoRA，这是一种用于稳定扩散（SD）的通用免训练加速模块。LCM-LoRA 可以作为独立且高效的基于神经网络的求解器模块来预测 PF-ODE 的解，从而能够在各种微调的 SD 模型和 SD LoRA 上以最少的步骤进行快速推理。文本到图像生成的大量实验证明了 LCM-LoRA 强大的泛化能力和优越性。

#### 5°C 贡献 & A 知识

这项工作建立在 Simian Luo 和 Yiqin Tan 的潜在一致性模型 (LCM) 的基础上 (Luo 等人, 2023)。基于 LCM，Simian Luo 编写了原始的 LCM-SDXL 蒸馏代码，并一起

与Yiqin Tan一起主要完成了这份技术报告。谭益勤发现了LCM参数的算术性质。Suraj Patil首先完成了LCM-LoRA的训练，发现其强大的泛化能力，并进行了大部分训练。Suraj Patil和Daniel Gu对原始LCM-SDXL代码库进行了出色的重构，提高了训练效率，并将其无缝集成到Diffusers库中。Patrick von Platen修订并完善了这份技术报告，并将LCM集成到Diffusers库中。黄龙波、李健、赵航共同顾问了原LCMs论文，并完善了本技术报告。我们还感谢Apolinário Passos和Patrick von Platen所做的出色的LCM演示和部署。我们还要感谢Sayak Paul和Pedro Cuenca帮助编写文档，以及Radamés Ajna创建演示。我们感谢Hugging Face Diffusers团队提供的计算资源来支持我们的实验。最后，我们非常重视LCM社区成员富有洞察力的讨论。

## 参考文献

邓家、董卫、理查德·索彻、李丽佳、李凯和李飞飞。Imagenet：一个大规模的高分层图像数据库。在2009年IEEE计算机视觉与模式识别会议，第248–255页。电子，2009。

乔纳森·何和蒂姆·萨利曼斯。无分类器的扩散指导。arXiv预印本 arXiv:2207.12598, 2022年。

尼尔·霍尔斯比、安德烈·朱吉乌、斯坦尼斯瓦夫·亚斯特热布斯基、布鲁纳·莫罗内、昆汀·德拉鲁西赫、安德雷亚·格斯蒙多、莫娜·阿塔里扬和西尔万·盖利。nlp的参数高效迁移学习。在国际机器学习会议，第2790–2799页。PMLR, 2019。

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, 还有陈伟柱。Lora：大型语言模型的低阶适应。arXiv预印本 arXiv:2106.09685, 2021年。

加布里埃尔·伊尔哈科、马可·图利奥·里贝罗、米切尔·沃茨曼、苏钦·古鲁兰甘、路德维希·施密特、汉纳内·哈吉什尔兹和阿里·法哈迪。使用任务算术编辑模型。arXiv预印本 arXiv:2212.04089, 2022年。

程璐、周宇豪、包凡、陈剑飞、李崇轩和朱军。Dpm求解器：A用于扩散概率模型采样的快速ode求解器，大约需要10个步骤。arXiv:2206.00927, 2022a。

程璐、周宇豪、包凡、陈剑飞、李崇轩和朱军。Dpm-solver++：快速用于扩散概率模型引导采样的求解器。arXiv预印本 arXiv:2211.01095, 2022b。

罗思勉、谭益勤、黄龙波、李健和赵航。潜在一致性模型：综合通过几步推理来调整高分辨率图像的大小。arXiv预印本 arXiv:2310.04378, 2023。

孟晨林、Robin Rombach、高瑞琪、Diederik Kingma、Stefano Ermon、Jonathan Ho等蒂姆·萨利曼斯。关于引导扩散模型的蒸馏。在IEEE/CVF计算机视觉和模式识别会议论文集，第14297–14306页，2023年。

吉列尔莫·奥尔蒂斯·希门尼斯、亚历山德罗·法维罗和帕斯卡·弗罗萨德。切线任务算术space：改进了预训练模型的编辑。arXiv预印本 arXiv:2305.12827, 2023。

达斯汀·波德尔、锡安·英格利希、凯尔·莱西、安德烈亚斯·布拉特曼、蒂姆·多克霍恩、乔纳斯·穆勒、乔佩纳和罗宾·罗姆巴赫。Sdxl：改进高分辨率图像合成的潜在扩散模型。arXiv预印本 arXiv:2307.01952, 2023。

罗宾·隆巴赫、安德烈亚斯·布拉特曼、多米尼克·洛伦茨、帕特里克·埃瑟和比约恩·奥默。高的-具有潜在扩散模型的分辨率图像合成。在IEEE/CVF计算机视觉和模式识别会议论文集，第10684–10695页，2022年。

克里斯托夫·舒曼、罗曼·博蒙特、理查德·文库、凯德·戈登、罗斯·怀特曼、迈赫迪 Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman 等。Laion-5b：用于训练下一代图像文本模型的开放大型数据集。*arXiv 预印本 arXiv:2210.08402*，2022 年。

分段思维。宣布 固态硬盘-1b: A 飞跃 在 高效的 t2i 一代。  
<https://blog.segmind.com/introducing-segmind-ssd-1b/>，2023 年。

宋嘉明、孟陈林和斯特凡诺·埃尔蒙。去噪扩散隐式模型。*arXiv 预印本 arXiv:2010.02502*，2020。

杨松、Prafulla Dhariwal、Mark Chen 和 Ilya Sutskever。一致性模型。*arXiv 预印本 arXiv: 2303.01469*，2023。

Fisher Yu、阿里·塞夫、张银达、宋舒然、托马斯·芬克豪瑟和肖建雄。孙：使用人类参与的深度学习构建大规模图像数据集。*arXiv 预印本 arXiv:1506.03365*，2015。

张静涵、陈世奇、刘俊腾、何俊贤。组成参数高效的模块与算术运算。*arXiv 预印本 arXiv:2306.14870*，2023。