结合强化学习与视觉生成模型：基础与进展

Yuanzhi Liang[1], Yijie Fang[1†], Rui Li[1†], Ziqi Ni[1†], Ruijie Su[1†], Chi Zhang[1], Xuelong Li[1*]

[1]Institute of Artificial Intelligence (TeleAI), China Telecom.

*Corresponding author(s). E-mail(s): xuelong_li@ieee.org;
Contributing authors: liangyzh18@outlook.com; fangyijie@stu.xidian.edu.cn;
rui.li@mail.ustc.edu.cn; zqni@seu.edu.cn; 472171770@qq.com;
zhangc120@chinatelecom.cn;
†These authors contributed equally to this work.

Abstract

Generative models have made significant progress in synthesizing visual content, including images, videos, and 3D/4D structures. However, they are typically trained with surrogate objectives such as likelihood or reconstruction loss, which often misalign with perceptual quality, semantic accuracy, or physical realism. Reinforcement learning (RL) offers a principled framework for optimizing non-differentiable, preference-driven, and temporally structured objectives. Recent advances demonstrate its effectiveness in enhancing controllability, consistency, and human alignment across generative tasks. This survey provides a systematic overview of RL-based methods for visual content generation. We review the evolution of RL from classical control to its role as a general-purpose optimization tool, and examine its integration into image, video, and 3D/4D generation. Across these domains, RL serves not only as a fine-tuning mechanism but also as a structural component for aligning generation with complex, high-level goals. We conclude with open challenges and future research directions at the intersection of RL and generative modeling.

Keywords: Reinforcement learning, diffusion models, generative models, image synthesis, video generation, 3D scene modeling, text-to-image, human feedback, multimodal learning.

1 介绍

最近在生成建模方面的进展——尤其是扩散模型 [1, 2] 和自回归方法 [3, 4] ——显著提高了图像 [5, 6] 、视频 [7, 8] 和 3D 内容合成 [9, 10] 的质量。尽管取得了这些进展，大多数生成模型仍然使用诸如最大似然估计或重构损失等替代目标进行训练，而这些目标往往与人类感知 [11, 12] 、语义 [13] 或物理合理性 [14] 不一致。因此，生成的输出可能会出现诸如运动不一致 [15, 16] 或结构错误 [17] 等瑕疵。

强化学习（RL）[18] 提供了一个通过互动优化不可微、符合人类需求或由环境驱动的目标的原理性框架。RL 最初是为控制和决策设计的，现已成为生成模型的强大补充，使得长时间跨度优化、反馈感知学习和基于偏好的微调成为可能。在最近的工作中，[5, 19, 20] ，RL 被越来越多地采用以改善可控性、增强现实性，并在多种模式下使生成与复杂且结构化的目标保持一致。

强化学习和生成建模的融合正在快速发展。如图 1 所示，处于这一交叉点的研究论文数量急剧增加——从 2019-2020 年的 13 篇增加到 2024-2025 年的 91 篇（截至 7 月 30 日）。仅在 2025 年上半年，就已经发表了 77 篇论文，预计今年将超过 140 篇出版物。这种指数式增长反映了从早期探索到更系统和广泛认可的研究范式的转变，这表明在更广泛的人工智能领域中，强化学习增强生成的重要性日益增加。

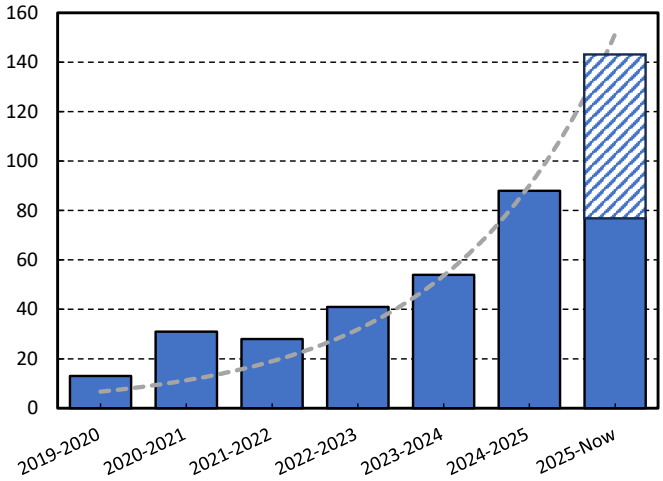

Fig. 1: 强化学习与视觉内容生成交叉领域的文献增长（2019-2025）。这个领域经历了指数级增长，论文数量从 2019-2020 年的 13 篇增加到 2024-2025 年的 91 篇（截至 7 月 30 日）。2025 年上半年已经发表了 77 篇论文，预计全年将超过 140 篇。这一趋势反映了该领域从探索阶段向巩固阶段的过渡，以及其在视觉生成研究中日益增长的战略重要性。

本综述全面概述了强化学习如何在多个领域中被整合到现代生成模型中。我们首先介绍了 RL 的概念演变——从其在解决马尔可夫决策过程中的起源，到其作为通用优化框架的崛起，嵌入于更广泛的学习系统中。接着，我们探讨了其在图像生成中的应用，其中 RL 技术通过策略优化和基于偏好的学习来提高语义对齐、美学质量和可控性。在视频领域，RL 有助于高效采样、时间一致性、物理可行性和与用户意图的对齐。最后，我们探讨了 RL 在 3D 内容生成中的应用，包括点云重建、网格合成、多视图一致性和人体运动建模。总之，这些脉络展示了 RL 作为一种多功能且具有原则性的机制，在引导生成模型朝向复杂、结构性且与人类对齐的目标方面发挥的日益重要的作用。

在这些领域中，强化学习充当了一个灵活且可解释的优化层，补充现有的架构。通过将生成重构为一个由评估反馈驱动的互动过程，强化学习使生成模型能够超越静态监督，向从偏好、结果或真实世界的约束中学习迈进。本文综述旨在描绘方法论格局、辨识跨域模式，并提供关于强化学习与生成模型交叉点上未来研究方向的见解。

2 RL 革命

强化学习（RL）[18] 提供了一个序列决策的通用框架，其中，智能体通过与环境交互来学习行动，以最大化长期奖励。其独特之处在于延迟反馈以及探索与利用之间的权衡。在过去十年中，RL 在仿真领域如电子游戏 [21, 22] 、机器人技术 [23–25] 和连续控制 [26, 27] 中取得了显著成功，展示了在定义明确的动态和奖励结构下的潜力 [21, 28, 29] 。然而，现实世界的应用仍然有限。实际部署常常存在样本效率低下 [30, 31] 、奖励指定的挑战 [32, 33] 以及泛化性差 [34, 35] 等问题。现实世界环境的多样性 [34, 36] 进一步复杂化了普遍有效算法的开发。

为了更好地理解强化学习（RL）不断发展的角色，本次综述采用了更广泛的视角：我们并不将 RL 视为一组固定的算法，而是将其框架为一种从交互、反馈和偏好中学习的一般范式。我们将这一演变组织为四个阶段，最终在其与生成建模和人类对齐学习的融合中达到高潮。

阶段 I: 通过试错解决马尔可夫决策过程 (MDPs) 的强化学习 [37, 38] [18, 39] 。

第二阶段：拆分为子领域（例如，离线 RL [40, 41] ，基于模型的 RL [42, 43] ）以克服特定的瓶颈。

阶段 III：对强化学习核心问题的重新定义——从已知环境中的策略优化转向学习环境动态 [44] 并与人类意图对齐 [33, 45] 。

阶段四：RL 作为决策的通用基质的出现，与规划 [46] 、模拟 [47] 和生成建模 [7, 19] 进行接口。

2.1 强化学习作为 MDP 求解器

强化学习的第一阶段基于马尔可夫决策过程 (MDPs) 的形式化，该过程通过一组状态、动作、转移动态和奖励函数来定义不确定性下的序列决策 [37, 38, 48] 。目标是计算一个最大化期望累积分数的最优策略。早期的强化学习努力与 MDP 的动态规划解决方案非常相似，假设对环境有完全的或抽样的访问。

两种主要的解决范式出现了：基于价值的方法和基于策略的方法。基于价值的方法通过评估操作或状态的期望回报（价值），并通过对这些估计值进行贪婪操作来导出策略。相反，基于策略的方法直接参数化并优化策略本身，通常通过期望奖励的梯度上升来实现。

价值基础的方法在历史上是主导的。时序差分（TD）学习 [49] 通过更新来自采样转换的价值估计，为蒙特卡罗评估与动态规划相结合奠定了基础。像 Q-learning [50] 和 SARSA [51] 这样的算法引入了异策略和同策略学习的变体，而对于小规模问题来说，表格实现是标准。

基于策略的方法因其对随机策略和连续动作空间的自然处理而获得了关注。REINFORCE 算法 [52] 引入了使用似然比技巧的蒙特卡罗策略梯度，尽管存在高方差。这导致了演员-评论家架构的产生，它通过一个学习的价值函数（评论家）来减少策略更新的方差。

深度学习的整合极大地拓展了这两种范式的应用范围。深度 Q 网络（DQN） [21] 利用卷积神经网络近似 Q 函数，并通过原始像素输入在 Atari 游戏中达到了人类水平的表现。后来，如 Double DQN [53] 、Dueling networks [54] 和优先级回放 [55] 等关键增强在 Rainbow 代理中被统一 [56] 。同时，具有可扩展性的基于策略的方法也出现了。异步优势演员评论（A3C）[57] 引入了并行处理和优势估计；信任区域策略优化（TRPO）[58] 及其实用变体近端策略优化（PPO） [59] 通过理论约束稳定了更新，成为连续控制中的标准。

深度学习的整合极大地扩展了两种范式的影响范围。深度 Q 网络（DQN）[21] 通过卷积神经网络来逼近 Q 函数，并利用原始像素输入在 Atari 游戏上达到了人类水平的表现。关键的增强方法如双 DQN [53] 、对抗网络 [54] 和优先级回放 [55] 后来在 Rainbow 代理 [56] 中得到了统一。与此同时，规模化的基于策略的方法也出现了。异步优势演员-评论者（A3C）[60] 引入并行化和优势估计；信任域策略优化（TRPO）[58] 及其实用变体近端策略优化（PPO）[61] 通过理论基础的约束来稳定更新，成为连续控制中的标准方法。

这些进展巩固了深度强化学习作为一种解决复杂马尔可夫决策过程（MDP）的强大工具，其在游戏中 [62] 、机器人操作 [63] 和模拟控制 [28] 的显著成功。然而，基于 MDP 的范式揭示了一些塑造下一阶段研究的基本限制。(1) 样本效率低下。大多数算法需要数百万次环境交互才能学习有效的策略——在仿真中可行，但在机器人或医疗等领域却过于繁琐。这种低效率源于数据重用性差、策略更新缓慢，尤其是低效的探索，在稀疏奖励场景中，代理难以发现有意义的轨迹或在长时间跨度上分配奖励。(2) 泛化能力有限。强化学习代理通常在与其被训练相同的环境中进行评估，导致对视觉或结构伪像的过拟合。像 Procgen 这样的基准显示，即便是微小的扰动也能显著降低性能，突显了鲁棒性差和传递能力弱的问题。(3) 不现实的假设。经典的强化学习方法通常依赖于完全可观察、静态的环境，具有明确的奖励和不受限制的交互，而这些条件在实际中很少能被满足。在许多真实世界任务中，交互是昂贵或不安全的，数据是离线或有限的，目标是隐含的或多方面的，这些假设显得不堪一击。

这些限制促使研究重心从解决固定的 MDP 转向放宽其假设—这导致了离线 RL、基于模型的方法以及对齐驱动的学习的兴起。这个转变标志着第二阶段的开始。

2.2 专业子领域的出现

离线强化学习、基于模型的强化学习以及其他子领域（多智能体 RL、安全 RL 等）作为半独立的研究领域出现。每个领域都引入了新的假设或额外的组件到 RL 范式中，以解决特定挑战：无需主动环境交互、整合动态知识、处理多个智能体或风险等。这种分化代表了一种范式转变：强化学习不仅仅被视为一个智能体解决 MDP 的问题，而是可以重新构建以适应数据可用性、环境知识或任务结构的限制。

离线强化学习：离线强化学习解决了从固定数据集中学习有效策略而无需进一步与环境交互的挑战。在探索代价高昂、风险高或无法实现的领域（如自动驾驶、医疗保健和工业控制）中，这种设置特别有价值。通过将强化学习重构为一个静态数据学习问题，离线强化学习使其更接近于监督学习，从而在实际系统部署中实现更具可扩展性和更高的安全性。

然而，这种范式引入了一个核心挑战：收集数据的行为策略与正在学习的目标策略之间的分布偏移。在无法访问环境的情况下，智能体可能会对分布外的动作赋予错误地高价值，导致外推误差和不稳定的训练 [40] 。最近的方法通过保守的价值估计或策略正则化解决了这一问题。例如，CQL [40] 惩罚未见过的动作的 Q 值以避免过高估计，而 BRAC [64] 和 TD3+BC [65] 约束学习策略以保持接近数据分布。IQL [66] 避免了显性约束，但修改了 Bellman 更新，以隐含地偏向于分布内的动作。这些进展，加上对自举误差和悲观主义的理论理解的增长，使得离线 RL 成为从静态日志安全、高效样本学习的一个有前途的解决方案，同时也在数据覆盖、评估和泛化方面带来了新的挑战。

多智能体强化学习：当多个智能体在同一环境中交互时，单智能体 MDP 框架不再有效，因为从每个智能体的角度来看，环境变得不固定性。这一挑战导致了多智能体强化学习（MARL） [67–70] 的发展，它解决了不固定性、合作、竞争和均衡策略等问题。关键的贡献包括 MADDPG [71] ，引入了集中训练和分散执行，这使得智能体在训练期间可以根据全局信息调整其评论者，而在测试时可以独立执行。其他方法，如 LOLA [72] ，通过预测一个智能体的更新如何影响其他智能体的学

习，纳入了对对手的建模。MARL 将经典的 RL 框架扩展到随机游戏和多智能体 MDP，利用博弈论的概念开发出稳定的学习动态和策略解决方案。

风险敏感和鲁棒强化学习：传统强化学习优化期望回报，可能忽视性能变化或安全关键故障。风险敏感方法 [73] 通过优化替代标准，如最小化方差或确保有利的最坏情况结果来解决这个问题。分布式强化学习模型 [74, 75] 整个回报分布而不是其平均值，使得决策更加丰富。鲁棒强化学习 [76–78] 进一步解决了动态或奖励中的不确定性，寻求在扰动或对抗性变化下表现良好的策略。技术如域随机化 [79] 和对抗性训练 [78] 已被用于增强鲁棒性并压力测试策略。这些方法反映了一种从最大化奖励向更加可靠且有韧性的目标的转变。

安全强化学习：安全强化学习专注于在训练和部署过程中满足显式的安全约束。它通常将问题建模为受约束的 MDP，引入如拉格朗日优化 [80] 、奖励塑形 [81] 或安全屏障 [82, 83] 的方法来限制不安全行为。这在自动驾驶 [84] 和机器人学 [85] 等领域尤其重要，因为不安全的探索可能导致灾难性的失败。安全强化学习也与探索研究相交，旨在开发在安全界限内仍能促进策略改进的学习策略。

在 2010 年代后期到 2020 年代中期，强化学习从优化已知环境中的固定奖励函数发展到作为更广泛学习系统中的模块化组件。这一阶段强调了两个关键方向：首先，使用强化学习来优化从数据中学习的模型，如模拟器或奖励函数；其次，使代理行为与无法通过标量奖励轻易捕捉的细致入微的人类目标对齐。焦点从改进策略优化算法转向设计学习任务本身的外部结构，包括如何建模偏好以及如何推断环境动态。强化学习逐渐由优化器转向问题陈述的方式。

从人类反馈中进行强化学习：从人类反馈中进行强化学习（RLHF） [33, 45] 已成为强化学习的一个高影响力应用，尤其是在使大型语言模型与人类意图对齐方面。RLHF 不是依赖于手工设计的奖励函数，而是使用人类偏好来定义目标。典型的流程包括收集模型输出的人类演示或排名 [33, 86] ，训练一个奖励模型来预测这些偏好，然后使用 RL 微调模型以最大化这个学习到的奖励 [45] 。这将诸如“提供帮助”这样的模糊目标转变为具体的 RL 问题：模型生成响应，奖励由偏好模型给出，而 RL 进行相应的优化。

InstructGPT [45] 的成功展示了这种方法的有效性。一个通过 RLHF 微调的 1.3B 参数模型在指令遵循上优于原始的 175B GPT-3，同时减少了有害或不相关的输出。RLHF 建立在早期的工作之上，比如以偏好为基础的 RL [33] ，该工作显示可以通过人类比较而非手动指定的方式学习奖励函数。在语言或视觉等高维空间中，这种反馈提供了可扩展的监督信号。重要的是，RLHF 的核心挑战不在于 RL 算法本身，而在于从主观数据中设计出稳健的奖励模型 [87] 。这一范式已经扩展到图像生成和机器人等领域，在这些领域中，人类对齐行为至关重要 [86] 。总的来说，RLHF 体现了对强化学习广义的看法——作为一个通过交互和评价性反馈将复杂模型与隐性目标对齐的灵活框架。

世界模型和基于模型的强化学习的复兴：基于模型的强化学习重新成为解决传统 RL 样本效率低下的一种有前途的方法 [42, 47] 。通过从数据中学习环境动态的模型，智能体可以在内部模拟交互，从而显著减少现实世界的交互进行策略训练。早期的工作如世界模型和 PETS 展示了在学习模拟器中训练的策略能够成功转移到真实环境 [42, 47] ，这突显了基于模型方法的潜力。

这一方向由 Dreamer 及其后继者进一步推进，后者学习隐空间动态模型并在设想的轨迹中完全训练代理 [43, 88] 。DreamerV2 [88] 尤其在使用更少环境帧的情况下，在 Atari 基准上实现了竞争性性能，展示了其可扩展性至高维、基于视觉的任务 [88] 。MuZero 采用了不同的方法，通过学习支持树搜索规划的隐式模型，将学习与规划结合而无需完整的转换模型 [44] 。

最近的架构，如 DreamerV3 和 Diffusion-QL，将世界模型与生成目标和偏好感知目标相结合，显示出从算法创新到任务规范和系统级设计的关注点转移 [89, 90] 。在这种观点下，强化学习越来越多地涉及首先学习潜在的马尔可夫决策过程——其动态和奖励——然后在该学习模型内进行规划或策略优化。这种两阶段的表述提供了更大的灵活性和数据效率，即使在学习的模型不完美的情况下，也已被证明是有效的。结果是强化学习的一个更宽广、更适应性的愿景，其中建模、模拟和优化深度交织在一起。

2.3 强化学习作为决策制定的一般基础

最近，人们越来越多地将强化学习视为一个通用的决策框架，而非一种独立的技术，适用于各种不同的领域。现代强化学习不仅关注固定环境中的交互代理，还嵌入到更大的系统中，包括规划、仿真、监督学习和人类反馈。

这个阶段的一个关键特征是在问题的制定上比算法的新颖性更强调。在像 RLHF [45] 这样的应用中，RL 算法的选择（例如，PPO [91] ）是次要的；关键部分是从人类偏好中学习到的奖励模型。类似地，在像 AlphaGo [92] 这样的系统中，RL 作为一个模块存在于包括模仿、自我对弈和树搜索的广泛架构中。在这些情况下，RL 作为一个微调层以优化一个明确的目标。

这种转变反映了一种更广泛的趋势：强化学习现在被视为一种通用的决策优化工具。它能够从隐性的或不可微的目标中学习，促进与复杂偏好的对齐，并扩展到结构化预测和生成建模。传统上使用监督学习解决的任务通常可以在 RL 术语中重新构思，其中奖励表示任务特定的指标。

随着强化学习更深入地与其他学习范式结合，其角色变得更加抽象且更加核心。它越来越多地被用于调节系统行为、指导架构搜索，甚至协调自我改进。与其将强化学习视为局限于交互环境的方法，第四阶段将其定位为一种统一的基础——能够在跨模态和目标的系统中驱动目标导向行为。

3 图像生成的强化学习

强化学习（RL）通过优化奖励信号，为图像生成模型提供了一个有效的指导框架，以实现期望的结果。与纯粹的监督方法不同，RL 在结合人类偏好以及保持输入（例如，提示或标题）与生成图像之间语义一致性方面表现优异。当前基于 RL 的图像生成方法主要分为两大类。第一类包括基于策略的方法，如近端策略优化（PPO）和最近的变体如组相对策略优化（GRPO），它们通过基于梯度的更新来优化策略。第二类是直接偏好优化（DPO），它将问题重新表述为使用已排序输出对的偏好分类任务。

Table 1: 按时间顺序编排的用于图像生成的强化学习方法的综合概述。

Method	RL Approach	Publication Date	Venue
DPOK[93]	Policy Gradient	May 2023	NeurIPS 2023
Promptist[94]	Policy Gradient	Dec 2022	NeurIPS 2023
DDPO[95]	Policy Gradient	May 2023	Arxiv
RLD[96]	Policy Gradient	May 2023	CVPR 2024
Diffusion-DPO[97]	DPO	Nov 2023	CVPR 2024
PRDP[98]	DPO	Feb 2024	CVPR 2024
Dense_Reward_T2I[99]	DPO	Feb 2024	ICML 2024
POSI[100]	Policy Gradient	Feb 2024	Arxiv
AGFSync[101]	DPO	Mar 2024	Arxiv
PAE[102]	Policy Gradient	Apr 2024	CVPR 2024
CurriculumDPO[103]	DPO	May 2024	CVPR 2025
HG-DPO[104]	DPO	May 2024	CVPR 2025
SPO[105]	DPO	Sat 2024	Arxiv
ReNO[106]	DPO	Jun 2024	NeurIPS 2024
DUO[107]	DPO	Jul 2024	NeurIPS 2024
RankDPO[108]	DPO	Oct 2024	Arxiv
PatchDPO[109]	DPO	Dec 2024	CVPR 2025
PPD[110]	DPO	Jan 2025	CVPR 2025
ImageGeneration-CoT[111]	DPO	Jan 2025	Arxiv
CaPO[112]	DPO	Feb 2025	CVPR 2025
DesignDiffusion[113]	DPO	Mar 2025	CVPR 2025
LightGen[114]	DPO	Mar 2025	Arxiv
SimpleAR[115]	GRPO	Apr 2025	Arxiv
DanceGRPO[116]	GRPO	May 2025	Arxiv
Flow-GRPO[117]	GRPO	May 2025	Arxiv
T2I-R1[118]	GRPO	May 2025	Arxiv
ReasonGen-R1[119]	GRPO	May 2025	Arxiv

基于

3.1 基于 PPO 的图像生成

PPO 的图像生成代表了在图像合成增强方面探索强化学习的早期尝试，特别是在使模型输出符合人类偏好、增强语义一致性以及实现可控生成。这些方法将图像生成，尤其是在扩散或自回归模型中，视为一个顺序决策过程，并利用学习的奖励信号来指导策略优化。

一个具有代表性的方法是去噪扩散策略优化（DDPO） [95] ，其将扩散模型中的去噪过程构建为一个多步 MDP。这使得模型可以直接优化下游生成目标，而不仅仅依赖于基于似然的训练。同时，DPOK [93] 也在这一方向上取得了进展，通过将文本到图像扩散模型的微调框定为多步马尔可夫决策过程（MDP）框架内的一个在线强化学习问题。它使用策略梯度优化反馈训练的奖励，关键是结合 Kullback–Leibler（KL）正则化以同时增强图像质量和文本-图像对齐。

为了促进输出多样性，Wallace 等人引入了基于最大均值差异 (MMD) 和互信息的奖励函数。这些奖励评估生成图像相对于参考集的分布覆盖率。每幅图像都会接收到基于边际效用的反馈，鼓励模型在保持样本质量的同时产生多样化的输出。

除生成阶段的优化外，基于 PPO 的方法也被应用于提示工程。Promptist [94] 提出了一个两阶段的流程，结合了监督微调（SFT）和基于 PPO 的强化学习，以自动优化用户提示。这改善了用户意图和模型内部偏好信号之间的匹配。POSI [100] 将此想法扩展到了安全的文本到图像生成，使用

自定义的奖励函数来平衡语义匹配和安全限制。PAE [102] 进一步引入了动态提示优化，实现了对提示权重和注入时机的细粒度控制。通过将 SFT 与在线 PPO 集成，PAE 在一个统一的奖励框架下优化美学、语义保真度和用户偏好。

3.2 基于 DPO 的图像生成

直接偏好优化（DPO）已经成为一种非常有效的方法，用于使图像生成模型与人类偏好对齐。早期的工作将 DPO 应用于标准生成架构。例如，扩散-DPO [97] 和 PRDP [98] 通过后训练优化将扩散模型与偏好数据对齐，提高了输出质量和可控性。

DPO 也被扩展到特定子任务。HG-DPO [104] 通过使用高质量的真实图像作为参考偏好，解决了人体图像生成中的结构和姿势不准确问题。RPO 引入了谐波奖励和提早停止，以提高主题导向生成，同时减少过拟合。PatchDPO [109] 集成了全局水平反馈以实现个性化生成，增强了局部一致性和细节保真度。类似地，DesignDiffusion [113] 应用了 DPO 从文本提示生成高质量的设计导向图像。

除了针对特定任务的改进外，还有一些工作将 DPO 应用于文本到图像生成中的更广泛挑战。ReNO [106] 通过在推理过程中优化初始噪声来改善组合细节，帮助模型在超出精调的情况下进行泛化。DUO [107] 将内容安全重新构建为偏好优化任务，旨在抑制不安全的输出。SPO [105] 通过基于偏好信号优化视觉吸引力，专注于美学对齐。

一些工作重新关注于增强 DPO 框架本身。AGFSync [101] 通过 AI 模型生成偏好注释，减少对人工标签的依赖。RankDPO [108] 和 CaPO [112] 提出可扩展的建模技术，以解决老化或昂贵的人类偏好数据的局限性。Dense Reward [99] 引入密集奖励信号，以捕捉细粒度的人类判断。Curriculum-DPO [103] 采用课程学习方法，通过逐步增加任务难度来提高样本效率。PPD [110] 将 DPO 扩展到建模个体用户偏好，解决群体级监督的局限性。

最后，DPO 也被应用于自回归模型。LightGen [114] 使用后期训练的 DPO 来提升生成质量。Image-Generation-CoT [111] 将 DPO 与链式思维 (CoT) 提示相结合，使得在序列生成任务中实现偏好对齐。这些努力反映了 DPO 在将图像生成模型与细致、以用户为中心的目标对齐方面的日益实用性和灵活性。

3.3 基于 GRPO 的图像生成

广义增强政策优化（GRPO）是政策优化的一个最新拓展，提供了改进的泛化能力和训练稳定性。作为 PPO 的一个变体，GRPO 在视觉内容生成中获得了关注，并被应用于扩散及流动基础的工作流程。

DanceGRPO [116] 提出一个统一的 GRPO 框架，支持多种生成范式，包括扩散模型和修正流。它适用于多项任务，如文本到图像、文本到视频和图像到视频生成。DanceGRPO 支持各种主干模型（例如，Stable Diffusion，FLUX）和奖励类型（例如，美学、一致性、运动），展示了 GRPO 在单一策略训练方案下统一异构任务和架构的灵活性。

Flow-GRPO [117] 将 GRPO 扩展到基于流的生成模型，通过将生成重新表述为一个随机微分方程 (SDE)。这使得基于强化学习的有效探索成为可能。它还引入了一个去噪步骤缩减策略，以提高样本效率和输出保真度。

GRPO 也已应用于自回归生成。T2I-R1 [118] 介绍了 BiCoT-GRPO，一个联合优化框架，它在单个 GRPO 循环内对齐语义级和令牌级的链式思维（CoT）推理。SimpleAR [115] 表明，GRPO 结合监督微调时，可以显著提升提示-图像的对齐和视觉质量。ReasonGen-R1 [119] 进一步将 CoT 推理整合到 GRPO 训练中，旨在改善跨多步生成过程的逻辑连贯性。

4 用于视频生成的强化学习

生成建模的最新进展，特别是在扩散和自回归架构方面，显著提升了从文本生成视频的质量。尽管取得了这些进步，大多数模型仍使用替代目标进行训练，例如最大似然估计（MLE）或重建损失。这些目标往往与人类观察者所重视的感知、语义或时间标准不一致。因此，生成的视频可能会出现例如运动不一致、语义漂移或物理上不合理的动态等瑕疵。

强化学习（RL）提供了一个有原则的框架来解决这些限制。它通过交互能够直接优化不可微分或偏好对齐的目标。在视频生成中，RL 已被用于结合结构化反馈、施加强制的时间和物理一致性，以及适应复杂的用户意图。

本节回顾了强化学习（RL）与视频生成交叉领域的最新进展。重点介绍了 RL 如何帮助提高生成视频系统的可控性、现实感和一致性。表格 **??** 总结了用于视频生成的代表性 RL 方法，展示了该领域从传统的策略优化到偏好驱动技术的演变。

一个实际应用于视频生成中的强化学习方法是优化扩散采样过程，以在不牺牲输出质量的情况下降低计算成本。AdaDiff 提议采用一种自适应步长选择方法，通过学习动态调整去噪步长来加速采

样。该过程被构建为一个强化学习问题，其中状态是当前的噪声视频帧和时间步长，动作是下一个去噪步骤的大小，奖励反映了生成速度和最终输出质量。训练一个策略网络来选择步长以平衡效率和保真度。当粗略更新足够时，选择较大的步长，而当需要细节时，则使用较小的步长。这种自适应调度使更快速的高质量视频生成成为可能，且所需去噪迭代次数更少。

4.1 顺序控制：规划和编辑

除了低级采样之外，强化学习（RL）通过将视频创作表述为一个序列决策任务，来实现高级程序控制。这包括在生成之前对内容和运动的规划，以及对现有视频的编辑或修整，以满足结构化和目标驱动的目标。

在视频规划中，RL 引导生成任务特定的目标。FLIP [120] 将指导性视频生成表述为一个剪辑选择问题。它采用演员—评论家框架，在该框架中，策略选择视频片段以实现文本指令。视觉-语言模型提供密集反馈作为奖励信号，使系统能够学习语义上对齐且视觉上连贯的视频计划。VideoAgent [121] 从具身 AI 视角进行规划。它利用基于演示的视频来启动视觉规划器，随后通过在实际环境中执行模拟过程来进行优化。尽管 RL 是学习循环的基础，该方法避免了显式策略梯度，而是使用奖励引导的数据选择。

除了规划，RL 还应用于视频编辑和转换。RL-V2V-GAN [122] 执行无监督的视频到视频翻译，例如，将夏季场景转换为冬季场景。它结合了 GAN 框架中的策略梯度，通过奖励输出的一致时间结构和风格保真度，使模型能够在没有配对训练数据的情况下学习一致的转换。E-Motion [123] 专注于使用事件相机输入进行未来运动预测。它将扩散的逆过程框架化为 MDP，并应用 PPO 以感知奖励（如 FVD 和 SSIM）来微调运动轨迹。

RL 也已经扩展到通用视频编辑领域。RLAVE [124] 提出了一个框架，其中角色选择编辑操作——如剪辑、转场或片段选择，而评论者使用视觉-语言模型来评估叙述连贯性、节奏和美感。该系统因制作风格一致且引人入胜的视频编辑而得到奖励，从而弥合自动合成与人工指导后期制作之间的差距。

4.2 与人类偏好的对齐

强化学习 (RL) 在视频生成中最具影响力的应用之一是后训练对齐。此阶段微调预训练生成器以更好地符合人类的主观偏好。对齐方法在如何整合奖励信号方面有所不同——从显式策略优化到基于梯度的偏好建模。

策略优化：一个核心方法群体使用正式的策略优化来对齐生成模型。虽然早期的强化学习算法在稳定性上存在挑战，但最近的方法提供了更稳健的性能。群体相对策略优化（GRPO）就是这样的一种方法。DanceGRPO [19] 将 GRPO 应用于各种视觉生成任务，包括文本到图像、文本到视频以及图像到视频的合成。它将扩散采样和校正流都重新定义为随机微分方程（SDEs），使得 GRPO 可以在不同架构和训练范式中操作。它支持多种基础模型（如稳定扩散、HunyuanVideo）和多种奖励类型，如美学质量、对齐、运动一致性和二进制反馈。这种灵活性使 GRPO 能够在后训练阶段有效地灌输结构化知识。

直接偏好优化：直接偏好优化（DPO）已成为将视频生成与人类偏好对齐的主流框架。与人类反馈强化学习（RLHF）相比，DPO 更稳定且更高效。它通过对静态的排序样本对数据集进行操作来跳过探索，并避免显式奖励建模。一个简单的分类损失促使模型复制优选的输出。

VideoDPO [125] 是一个早期的实现，使用自动评分启发式方法构建大规模偏好数据集。HuViDPO [126] 将该框架扩展到文本到视频生成。它提出了一种结构化损失和架构改进方案，例如首帧条件和稀疏因果注意力，以改善视觉质量和时间一致性。

为了支持细粒度的对齐，一些方法将偏好分解为可解释的维度。VisionReward [127] 使用多目标偏好优化（MPO）来平衡时间一致性、文本保真度和现实性。HALO [128] 进一步引入了 Granular-DPO，它应用了贴片级反馈来修正局部的瑕疵。DenseDPO [129] 通过对受损视频去噪并使用片段级注释进行监督来改善结构对齐。这在保持质量的同时降低了注释成本。

偏好数据的获取进展也推动了 DPO 的可扩展性。RDPO [7] 使用基于物理的启发式方法自动从真实视频中生成偏好对。例如，倾向于掉落的球而不是漂浮的球。这种方法不需要人工标记即可编码物理合理性。

一些研究改进了 DPO 本身的学习动态。OnlineVPO [130] 提出了一种在线变体，该变体实时地连续采样、排序和更新模型。这减少了时间伪影，提高了主体一致性，而不依赖于静态数据集。

此外，DPO 也已被整合到混合训练流程中。SePPO [131] 和 EchoMimicV3 [132] 在有监督学习和偏好优化之间交替进行，结合了有监督训练的稳定性和 DPO 的对齐精度。VPO [133] 扩展了这一理念，通过 DPO 优化用户提示而非生成模型。SkyReelsV2 [134] 实现了一个多阶段流程，涉及监督微调、运动特定的 DPO、无限长度生成和最终抛光。这种结构化框架支持开源电影级视频合成。

DPO 目标的新表述也已经出现。AAVG [135] 引入了基于差距感知的偏好优化（GAPO），其根据偏好信心权重训练样本。IDOI [136] 在统一的概率框架中结合了 DPO 和奖励加权回归（RWR），利用二元视觉语言反馈优化动态对象交互。IPO [137] 提出了一种基于 RL 的迭代偏好优化框架，在评估和评论指导更新之间交替进行。这改善了语义保真度、运动逼真度和美学吸引力。相比之下，Diffusion-NPO [138] 专注于不良输出。它引入了负偏好优化（NPO），通过倒置的偏好标签训练模型以抑制不需要的行为。负对齐模型在无分类器指导中用作无条件分支，增强了高质量和低质量生成之间的区分。这种策略改善了任务的视觉质量、鲁棒性和与人类的对齐。

奖励微调。尽管与经典的强化学习不同，奖励微调包含一系列使用奖励信号直接监督生成模型的方法。这些方法通常将奖励视为优化目标，绕过了策略学习和价值估计。虽然严格意义上不属于强化学习，但它们有一个共同的目标：使生成的输出与高级人类偏好对齐。

InstructVideo [139] 代表了这一范式的一个早期实例。它将视频生成设计为一个编辑任务，并利用预训练的图像级别奖励模型来评估输出。这种轻量策略通过迭代地应用奖励反馈而无需完全重新训练模型来减少计算开销。VADER [140] 通过将多个专家模型（例如用于语义对齐的 CLIP 和用于物体检测的 YOLO）整合到一个统一的可微分奖励函数中来扩展这一框架。这使得可以在多个质量维度上进行基于梯度的优化，从而提高输出的忠实度和相关性。

与这些基于梯度的方法相反，Free ^{2} Guide [141] 采用了一种受到最优控制启发的无梯度方法。它使用黑盒奖励信号——例如来自视觉-语言模型的评估——来引导推理时的决策。该即插即用策略根据奖励信号来调节生成过程，而无需修改生成器的参数。

4.3 物理一致性与世界建模

越来越多的研究探讨如何通过强化学习来增强视频生成中的物理现实感和世界建模。这些方法旨在产生不仅看起来合理，而且还尊重基础物理原则和环境动态的内容。

Phys-AR [142] 是这一方向上的一个开创性尝试。它在后训练阶段将符号推理与强化学习结合起来。模型首先使用扩散时间步分词器（DDT）将视频帧转换为符号化标记。然后，将生成过程框架化为标记级的 MDP，并应用 GRPO 来优化生成策略。Phys-AR 的一个关键贡献在于其奖励设计，其中编码了物理定律，如速度一致性和质量信息运动。这鼓励模型生成展示匀速运动、抛物线轨迹和逼真碰撞反应的运动轨迹。

RLVR-World [143] 提供了一种互补的方法，解决了视频生成中 MLE 的基本局限性。MLE 通常产生模糊或语义不稳定的输出，因为它没有针对感知质量进行优化。RLVR-World 引入了一种基于强化学习的微调方案，直接针对感知奖励进行优化，包括 LPIPS、SSIM 和任务准确性。GRPO 再次被用于使生成策略与这些指标对齐。重要的是，这个框架不仅限于视频，还展示了其在语言建模和其他生成任务中的适用性。

虽然 RL 通常在训练期间用于对齐或控制，最近的工作已经研究了其在推理时用于动态指导生成决策。这对于长篇视频合成尤其有用，因为模型必须在较长的序列中保持时间一致性、语义连贯性和提示忠实性。

InfLVG [144] 通过将 GRPO 直接整合到推理循环中展示了这一范例。不同于依赖静态的上下文选择，InfLVG 将生成过程公式化为一个顺序决策过程。在每一步，模型对一组候选延续进行采样，并通过一个综合奖励函数对它们进行评分。该奖励平衡三个关键方面：（1）通过面部身份相似性测量的内容一致性；（2）通过基于 CLIP 的文本-视频对齐评估的提示相关性；以及（3）使用视觉-语言分类器进行惩罚的伪影抑制。这些分数用于计算候选者之间的相对优势。GRPO 然后通过一个剪辑目标更新策略，指导 Plackett-Luce top-K 采样过程选择最合适的上下文标记。此推理时优化在保持计算预算固定的同时，实现了自适应上下文选择。结果是，InfLVG 在提高视觉清晰度和时间一致性的同时实现了显著的长度扩展——最长可延长至 9 倍。

5 用于 3D 内容生成的强化学习

将强化学习（RL）融入 3D 内容生成引入了一种强大的范式，结合了顺序决策和几何建模。这种协同作用使复杂的 3D 结构的合成更加自适应、可控且符合偏好。如图 2 所示，该领域的最新研究工作广泛分布于多个子领域，包括文本到 NeRF/3D 高斯溅射、3D 扩散建模、多视图一致性优化、人体运动生成、点云重建以及其他新兴方向。这些主题之间的平衡分布突显了 RL 在 3D 生成中的应用广度和日益成熟。与早期阶段以单一任务或表示为主导不同，当前趋势表明 RL 正在整个 3D 内容生成流程中被探索作为一种通用优化框架。

为了整理这越来越多的研究成果，我们将近期的进展分为三个核心领域：(1) 通用 3D 内容生成，(2) 多视角模型增强，(3) 特定领域应用，例如表面补全、立体场景重建和动态人体运动合成。

Table 2: 在选定的 VBench 子指标上的评估。

Models	Image Quality	Multiple Objects	Human Action	Motion Smoothness	Subject Consistency	Aesthetic Quality
VC2+VideoDPO [125]	-	52.29	99.00	92.18	95.69	63.18
Turbo+VideoDPO [125]	-	51.98	94.00	88.85	96.10	68.98
CogVid+VideoDPO [125]	-	54.04	81.00	88.64	94.67	58.64
VisionReward [127]	-	71.54	98.40	-	-	-
Turbo-v1+HALO [128]	72.07	54.97	95.00	-	-	-
Turbo-v2+HALO [128]	69.11	67.5	97.6	-	-	-
CogVidX-2B+HALO [128]	61.90	72.91	98.00	-	-	-
RDPO [7]	65.11	-	-	99.27	97.04	55.01
CogVidX-2B+VPO [133]	-	70.17	99.00	-	-	-
CogVidX-5B+VPO [133]	-	75.73	99.60	-	-	-
AAVG [135]	68.98	-	-	99.13	95.20	54.28
OnlineVPO [130]	67.36	-	-	99.36	97.58	55.37
InstructVideo [139]	70.09	-	-	96.76	96.45	50.01
VADER [140]	66.08	-	-	98.89	95.53	53.43
CogVidX-2B+IPO [137]	62.87	-	-	98.17	96.79	62.31

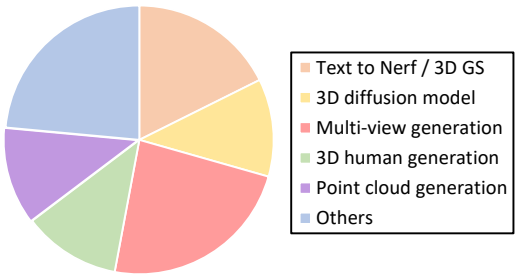

Fig. 2: 该图显示了最近将强化学习应用于 3D 生成领域的研究课题的比例分布，包括文本到 NeRF/3D 高斯喷溅、3D 扩散模型、多视图一致性优化、人体运动合成、点云建模等。各个领域之间的均衡分布反映了这是一个快速出现和多样化的领域，没有任何一个范式占据主导地位。这表明，RL 正被广泛地探索为用作整个 3D 生成任务光谱中的通用优化工具。

5.1 3D 内容生成

基于点云的 3D 生成：Akizuki 等人 [145] 是最早探索用于 3D 生成的 RL 方法之一。他们的方法将 3D 模型体素化以定义空间边界。动作空间对应于体素扩展的方向，而奖励则鼓励拓扑有效性并惩罚自相交。这种设置使得代理能够通过学习体素网格中的运动轨迹来生成多样且结构良好的 3D 对象。

最近的研究将这一理念扩展到网格生成中。DeepMesh [146] 在 3D 自回归模型中使用直接偏好优化（DPO），以使生成与人类偏好一致。然而，该方法依赖于昂贵的偏好对的人工注释。Mesh-RFT [147] 提出了一种拓扑感知评分方案，使用自动化指标——边界边缘比率（BER）和拓扑得分（TS）——在对象和面水平评估网格质量。它进一步引入了 M-DPO 和质量感知掩模，以改进有缺陷的区域并捕捉局部几何变化。

基于深度图像的 3D 生成：Zhang 等人 [148] 提出了一种基于 A3C 的深度强化学习方法，用于从 RGB-D 输入中重建完整的 3D 场景，即使在遮挡严重的情况下。该模型通过学习的策略选择最佳视点，并应用体积引导的修补进行渐进的语义补全。类似地，Lin 等人 [10] 使用深度图像作为结构参考，训练两个 RL 代理，一个用于修改原始对象，另一个用于调整顶点。他们的奖励基于交并比（IoU）的变化，从而实现精确的几何编辑。

文本和图像条件的 3D 生成：最近 3D 生成的进展越来越多地探索使用强化学习以提高可控性、几何一致性和人类偏好对齐。特别是，RL 提供了一个优化不可微目标的自然框架，使得能够细粒度地基于奖励来调整从文本或图像提示生成的 3D 资产。

DreamReward [9] 代表基于 RL 的框架，并构建了一个大规模偏好标记数据集，该数据集包含 25,304 个提示-3D 资产对及人类反馈分数。它训练一个奖励模型，Reward3D，以评估生成质量。采样过程被表述为一个 RL 任务，动作对应于 3D 表示的更新，而奖励则来源于 Reward3D。通过将这个学习到的奖励与 Score Distillation Sampling (SDS) [149] 整合，该方法改善了 3D 输出的对齐和保真度。

然而，DreamReward 对于专用奖励模型的依赖引入了可扩展性的限制。DreamDPO [150] 通过应用直接偏好优化（DPO）解决了这个问题，其利用大型预训练视觉-语言模型作为零样本奖励函数。

在训练过程中，通过文本指令对 3D 资产的图像渲染进行排序，这些排序指导了基于偏好的 3D 表示更新。该框架消除了对手动奖励监督的需求，同时利用了大型多模态模型的泛化能力。通过将过程框架化为离线偏好优化，DreamDPO 与强化学习原则对齐 3D 生成，同时避免了动态探索的不稳定性。

为减轻对偏好标记数据的需求，DreamCS [151] 提出了一种可扩展的替代方案，使用未配对的 3D 网格和基于柯西–施瓦茨（CS）散度的奖励目标。它在这个未配对的数据集上训练一个几何感知的奖励模型 RewardCS，并应用奖励引导采样来逐步改善 3D 输出。虽然没有明确地构建为强化学习代理-环境循环，但 DreamCS 仍通过定义以奖励为驱动的优化景观和结构化反馈来遵循 RL 范式。

近年来的方法也将强化学习（RL）整合到大规模三维扩散模型的训练中。GPLD3D [152] 构建了两个领域特定的奖励函数——几何可行性和物理稳定性——并将它们纳入生成循环，以指导高质量资产的合成。在一个更具普适性的方法中，Nabla-R2D3 [153] 提出了一种新颖的 RL 对齐框架，它将二维和三维奖励空间连接起来。它使用一个预训练的二维奖励模型通过概率精化机制 Nabla-GFlowNet 来指导三维生成，该机制将二维信号转化为结构化的三维奖励。这一方法突出了三维生成中强化学习的关键优势：即使在缺乏明确三维监督的情况下，也能利用代理信号跨模态进行优化。

5.2 多视图模型增强

除了直接生成 3D 资产之外，强化学习 (RL) 越来越多地被应用于提高多视角一致性，这是 3D 生成流程中的一个关键挑战，特别是由于 Janus 问题，其中合成物体的不同视角显示出冲突的几何形状或外观。

Carve3D [154] 将增强学习微调 (RLFT) 引入到多视点扩散模型中。Carve3D 从预训练的文本到图像模型开始，使用一种新颖的多视点重建一致性 (MRC) 指标进行 RLFT。该指标将从多个视点生成的图像与通过大型稀疏视点模型 (LRM) 重建的 NeRF 渲染视图进行比较。MRC 分数作为奖励信号，引导扩散模型生成在视点之间视觉上合理且几何上一致的输出。

HFDream [155] 在这一方向上更进一步，通过构建一个大规模的多视图图像–文本对数据集，这些对被标注了人类反馈。一个奖励模型被训练用来评估生成图像相对于文本提示的视点准确性。然后使用该模型通过奖励加权损失微调扩散生成器，将多视图输出在文本到 3D 场景中直接对齐到人类偏好和几何保真度。

进一步将强化学习集成推进到图像到 3D 领域，MVReward [156] 收集了一个包含多视角 RGB 图像、法线贴图以及相关的人类偏好注释的数据集。它引入了一种多视角偏好学习策略，该策略将从学习的偏好模型导出的奖励损失与标准预训练目标结合在一起。这种混合的强化学习公式提升了重建的 3D 形状的结构真实感和视点对齐度。

为了缓解少步文本到多视图模型中高计算成本与生成质量之间的矛盾，Zhang 等人 [157] 提出了一个新的强化学习微调框架，将任务重新表述为一个约束的单视图策略优化问题。他们的方法引入了 ZMV-Sampling，提高了跨视图和文本图像对齐以及 MV-ZigAL，一个奖励优势学习方案。MV-ZigAL 使用优势加权策略更新来内部化基于奖励的改进，使基础采样策略能够用更少的推理步骤生成连贯的多视图图像。

5.3 其他特定领域的 3D 应用

强化学习不仅被用来提高一般 3D 资产的质量和对齐，还被扩展到需要结构化控制、物理推理或多对象协调的特定领域生成任务。这些应用进一步展示了强化学习作为视觉内容生成的一般决策支持框架的多功能性。

QINet [158] 将强化学习应用于 3D 点云补全，使用一个 Actor-Critic 框架。代理与一个潜在空间 GAN 交互，优化噪声向量以将损坏的输入重建为完整的形状。奖励结合了几何准确性（IoU）和潜在一致性，即使在大范围缺失区域的情况下也能实现强大的补全能力。RLSS [159] 解决序列化的 3D 场景生成。它将场景构建形式化为一个强化学习问题，其中状态包括 2D 布局和物体放置，而动作则选择和放置物体。通过 PPO，代理接受奖励强化物理约束和设计目标的训练。该方法以最小监督实现多样且逼真的室内布局。赵等人 [160] 提出了一种场景感知的 RL 框架，用于在 3D 环境中生成人体动作。PPO 用于在潜在动作空间中训练策略。状态捕捉几何、身体姿态和交互目标，奖励鼓励目标完成、接触准确性和碰撞避免。该方法支持复杂的交互和细粒度的路径控制。Bailando [161] 引入了一种基于 transformer 的 RL 框架，用于 3D 舞蹈生成。编舞记忆编码量化的姿势，一个 GPT 风格的 Actor-Critic 预测未来动作。节拍对齐奖励确保与音乐同步以及身体部位之间的一致性，生成逼真且风格一致的舞蹈。Text2Stereo [162] 处理文本到立体图像生成。它通过 AlignProp 在垂直堆叠的立体对上微调 Stable Diffusion，并使用 RL。可微奖励引导优化以保持立体一致性和提示对齐，并将人类偏好整合到反馈信号中。

这些例子突出了强化学习在增强跨不同领域的 3D 生成中的多功能性。通过动态反馈和结构化控制，强化学习将生成模型扩展到更复杂和更互动的 3D 场景。

Table 3: 应用于 3D 生成的强化学习方法的全面概述。

Approaches	Year	Task	Training Techniques	Actions	Reward	States	Performance
Object Generation with Design Constraint [145]	2020 RE	Constraint-aware 3D Objects Generation	ML	Neighboring voxels (6 or 26 directions) moving or specific LEGO blocks placing, depending on the experiment	+1 (inside geometry), -0.5/-1 (self-intersection / outside)	Local voxel window (e.g., 5x5x5) around the agent	Generates topologically consistent, stylistically varied 3D objects and LEGO assemblies, topologically consistent designs meeting constraints
DeepMesh [146]	2025 Arxiv	Mesh Generation	DPO	Vertex/face token prediction in hierarchical blocks	Human preference + geometric metrics (Chamfer Distance)	Local voxel window (e.g., 5×5×5) + patch connectivity.	SOTA in geometric fidelity, visual appeal, generation efficien and face count (30k+)
Mesh-RFT [147]	2025 Arxiv	Mesh Generation	Masked DPO	Face-level token prediction with quality-aware masking	BER (Boundary Edge Ratio) + TS (Topology Score) + HD (Hausdorff Distance) metrics	Tokenized mesh sequences + local quality masks.	SOTA in both geometric accuracy and visual fidelity
Point Cloud Scene Completion from Single RGB-D Image [148]	2023 TPAMI	Colored Semantic Point Cloud Scene Completion	A3C	Next-best-view selection	Accuracy, hole-filling, and point cloud recovery	Updated point cloud at each iteration	SOTA in scene completion and scene segmentation accuracy
3D Shapes Modeling [10]	2020 ECCV	3D Shape Modeling	IL + RL	Primitive manipulation and mesh editing	IoU-based and parsimony rewards	Shape reference, primitives/edge loops, step encoding	Structurally meaningful and regular meshes
DreamReward [9]	2024 ECCV	Text-to-3d Generation	RLHF	Optimizing 3D representations using human feedback	Reward3D (human preference scoring)	3D representations	SOTA in high-fidelity, human-aligned 3D generation
DreamDPO [150]	2025 Arxiv	Text-to-3D Generation	DPO	Optimizing 3D representations via pairwise comparisons	Preference scores from reward/LMM models	3D representations	SOTA in high-quality, controllable 3D generation
DreamCS [151]	2025 Arxiv	Text-to-3D Generation	CS divergence	Optimizing 3D representations via unpaired preference learning	RewardCS (3D geometry-aware scoring)	3D representations	SOTA in high-fidelity, geometrically consistent 3D generation
Nabla-R2D3 [153]	2025 Arxiv	Text-to-3D Generation	Gradient-informed RL Finetune	Sampling and rendering multi-view images	2D appearance and geometry scores.	Latent 3D shape representations	Propose a method Nabla-R2D3 that can effectively, efficiently and robustly finetune 3D-native generative models from 2D reward models with better preference alignment, better text-object alignment and fewer geometric artifacts
Carve3d [154]	2024 CVPR	Text-to-Multiview Generation	On-policy Policy Gradient, KL regularization	Denoising steps in diffusion process	MRC (Multi-view Reconstruction Consistency) metric	Text prompt, timestep, noisy image	Superior multi-view consistency and NeTF reconstruction quality
HFDream [155]	2024 OpenReview	Text-to-Multiview Generation	LoRA, SFT	Generating view-aligned images	Human-annotated view alignment score	Text prompt and diffusion model output	Improved 3D consistency and text alignment
MVReward [156]	2025 AAAI	Image-to-Multiview Generation	MVP	Denoising steps in diffusion process	Human preference score (MVReward)	Image prompt and noisy multi-view images	Improved human-aligned 3D generation quality
Refining Few-Step Text-to-Multiview Diffusion [157]	2025 Arxiv	Text-to-Multiview Generation	MV-ZigAL, Constrained Optimization	Denoising steps in diffusion process	Joint-view scores including alignment quality, geometry quality, texture quality, and overall quality, and single-view fidelity scores	Multiview-aware MDP (Markov decision process) with text prompt and viewpoint embeddings	Improved fidelity and consistency while preserving few-step T2MV diffusion baseline efficiency
QINet [158]	2022 TGRS	Point Cloud Completion	DDPG	Revised noise of z (clean latent code)	Weighted sum of reconstruction and latent rewards	Incomplete representation in latent space (IRLS)	SOTA in shape completeness, perceptual quality, shape resolution and category distinction

RLSS [159]	2022 WACV	3D Indoor Scene Generation with Predefined Design Objects	PPO	Object placement and selection	Constraint-driven rewards that include multiple conditions such as successful condition, count of objects, and failure conditions	2D scene representation with object metadata	High-quality, diverse scene generation with fast synthesis
Synthesizing diverse human motions in 3d indoor scenes [160]	2023 ICCV	Human Motions Synthesis	PPO	Latent variables of CVAE-based generative motion primitive	Goal-reaching, foot-ground contact, and penetration avoidance	configurations of 3D scene geometry, virtual human body, and intended goals to interact with	Realistic and diverse human-scene interactions with low penetration, high foot-ground contact accuracy, and efficient task completion times, outperforming baselines in perceptual studies
Bailando [161]	2022 CVPR	3D Dance Generation	Actor-Critic Finetune	Predicting next dance pose	Beat-align reward and half-body consistency reward	A sequence of dance poses	SOTA in motion quality, diversity, and music-dance alignment

6 机制与见解

强化学习不仅仅提供了一种替代的优化策略——它引入了一种根本不同的视角，关于如何引导、调整和校准生成模型。在图像、视频和 3D/4D 视觉内容中，强化学习已被证明是一种多功能的机制，能解决传统监督目标难以捕捉的挑战。下面，我们总结了强化学习促进生成建模进步的核心机制，并提供对未来研究的关键见解和影响。

6.1 机制：强化学习的功能

非可微目标的优化：许多生成目标——例如美学吸引力、用户满意度或物理合理性——无法表达为可微损失。强化学习允许通过奖励信号、人类反馈或替代理评价直接优化这些目标。
细粒度控制和可控性：RL 天然适用于处理顺序和条件决策，使其非常适合指导细粒度生成过程。这允许对生成步骤、提示或结构进行动态控制，特别是在扩散、自回归或程序化模型中。
引入结构化和时间反馈：特别是在视频和 3D/4D 生成中，RL 利用序列反馈以确保时间一致性、因果关联和物理有效的轨迹——提供比静态配对数据更丰富的监督。
人类偏好对齐：通过偏好建模，RL 提供了一个原则性框架，将生成与主观目标和用户意图对齐。这使得模型能够内化隐含的、模糊的或任务特定的反馈。

6.2 见解：更广泛的意义

强化学习作为一种结构性组件，而不仅仅是训练技巧：强化学习不仅越来越被用作训练后的优化工具，还被用作生成系统设计的核心部分——支持奖励建模、基于策略的采样以及考虑反馈的调度。

偏好基础范式的崛起：DPO 及其扩展重新定义了 RL 与生成模型的互动方式，将重点从注重探索的训练转向稳定、可扩展的偏好对齐。这标志着以样本效率为基础和人类对齐生成的新纪元。
统一建模和交互：RL 弥合了生成学习和交互学习之间的差距。它为开发适应性强、有实体感和自我意识的生成代理提供了基础，这对于在创意、科学或机器人环境中的实际应用至关重要。

超越静态目标：迈向多目标、多智能体生成：随着强化学习的不断发展，我们预期将转向多目标优化，在这种情况下，生成模型必须在质量、多样性、安全性和效率之间权衡取舍——这可能是在多智能体或合作学习设置中进行的。

6.3 未来工作

强化学习和生成建模的融合正在改变我们对生成的思考方式——不是将其视作从输入到输出的静态映射，而是视为一个互动的、迭代的和目标驱动的过程。随着生成系统变得更加自主和面向用户，从反馈中学习、优化复杂目标和适应多样化偏好的能力将变得不可或缺。强化学习提供了理论工具和算法基础以满足这些需求。

7

缩写

- RL ： 强化学习
- MDP：马尔可夫决策过程
- PPO：近端策略优化
- GRPO：组相对策略优化
- DPO：直接偏好优化
- CFG：无分类器引导
- RLHF：带有人类反馈的强化学习
- VPO ： 视觉偏好优化
- IPO：迭代偏好优化
- NPO：负偏好优化
- LPIPS：学习感知图像块相似度
- 结构相似度指数
- GAN ： 生成对抗网络
- NeRF：神经辐射场
- GPT ： 生成型预训练变换器
- CLIP：对比语言–图像预训练
- BLIP-2：自举语言-图像预训练 v2
- YOLO：你只看一次
- A3C：异步优势演员-评论者

- IoU：交并比
- RGB-D：红-绿-蓝与深度
- BER：边界边比率
- 拓扑得分
- MV ： 多视图
- RLFT：强化学习微调
- MRC：多视图重建一致性
- ZMV：Zigzag 多视图（在 ZMV-Sampling 中）
- MV-ZigAL：多视角 Zigzag 优势学习
- DDT：扩散时间步标记器
- LRM：大型重建模型

References

[1] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.-H.: Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys **56**(4), 1–39 (2023)

[2] Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10850–10869 (2023)

[3] Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al.: Autoregressive models in vision: A survey. arXiv preprint arXiv:2411.05902 (2024)

[4] Kaur, J., Parmar, K.S., Singh, S.: Autoregressive models in environmental forecasting time series: a theoretical and application review. Environmental Science and Pollution Research **30**(8), 19617–19641 (2023)

[5] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)

[7] Qian, W., Wang, C., Peng, H., Tan, Z., Li, H., Zeng, A.: Rdpo: Real data preference optimization for physics consistency video generation. arXiv preprint arXiv:2506.18655 (2025)

[8] Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)

[9] Ye, J., Liu, F., Li, Q., Wang, Z., Wang, Y., Wang, X., Duan, Y., Zhu, J.: Dreamreward: Text-to-3d generation with human preference. In: European Conference on Computer Vision, pp. 259–276 (2024). Springer

[10] Lin, C., Fan, T., Wang, W., Nießner, M.: Modeling 3d shapes by reinforcement learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, pp. 545–561 (2020). Springer

[11] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)

[12] Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)

[13] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, S., Sutskever, I.: Learning transferable visual models from

natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[14] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)

[15] Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H.: Learning to generate long-term future via hierarchical prediction. In: International Conference on Machine Learning, pp. 3560–3569 (2017). PMLR

[16] Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)

[17] Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International Conference on Machine Learning, pp. 3481–3490 (2018). PMLR

[18] Sutton, R.S., Barto, A.G., *et al.*: Reinforcement Learning: An Introduction vol. 1. MIT press Cambridge, ??? (1998)

[19] Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)

[20] Zhou, D., Li, Y., Li, Q., Yang, Y., Tang, J., Shen, Y., Li, X., Wang, X., Zhou, P.: Flow-grpo: Training flow matching models via online reinforcement learning. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024). https://arxiv.org/abs/2312.06699

[21] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., *et al.*: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)

[22] Shao, K., Tang, Z., Zhu, Y., Li, N., Zhao, D.: A survey of deep reinforcement learning in video games. arXiv preprint arXiv:1912.10944 (2019)

[23] Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., Levine, S.: How to train your robot with deep reinforcement learning: lessons we have learned. The International Journal of Robotics Research **40**(4-5), 698–721 (2021)

[24] Zhao, W., Queralta, J.P., Westerlund, T.: Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 737–744 (2020). IEEE

[25] Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., *et al.*: Scalable deep reinforcement learning for vision-based robotic manipulation. In: Conference on Robot Learning, pp. 651–673 (2018). PMLR

[26] Dulac-Arnold, G., Levine, N., Mankowitz, D.J., Li, J., Paduraru, C., Gowal, S., Hester, T.: Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. Machine Learning **110**(9), 2419–2468 (2021)

[27] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)

[28] Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D.d.L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., et al.: Deepmind control suite. arXiv preprint arXiv:1801.00690 (2018)

[29] Wang, S., Dai, W., Li, G.Y.: Distributionally robust receive beamforming. arXiv preprint arXiv:2401.12345 (2024)

[30] Mohanty, S., Poonganam, J., Gaidon, A., Kolobov, A., Wulfe, B., Chakraborty, D.,

Šemetulskis, G., Schapke, J., Kubilius, J., Pašukonis, J., et al.: Measuring sample efficiency and generalization in reinforcement learning benchmarks: Neurips 2020 procgen benchmark. arXiv preprint arXiv:2103.15332 (2021)

[31] Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R.H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al.: Model-based reinforcement learning for atari. arXiv preprint arXiv:1903.00374 (2019)

[32] Ng, A.Y., Harada, D., Russell, S.: Policy invariance under reward transformations: Theory and application to reward shaping. In: Icml, vol. 99, pp. 278–287 (1999). Citeseer

[33] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)

[34] Zhang, C., Vinyals, O., Munos, R., Bengio, S.: A study on overfitting in deep reinforcement learning. arXiv preprint arXiv:1804.06893 (2018)

[35] Arjovsky, M.: Out of distribution generalization in machine learning. PhD thesis, New York University (2020)

[36] Dulac-Arnold, G., Mankowitz, D., Hester, T.: Challenges of real-world reinforcement learning. arXiv preprint arXiv:1904.12901 (2019)

[37] Bellman, R.E., Dreyfus, S.E.: Applied Dynamic Programming. Princeton university press, ??? (2015)

[38] Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, ??? (2014)

[39] Sutton, R.S.: Dyna, an integrated architecture for learning, planning, and reacting. In: SIGART Bulletin, vol. 2, pp. 160–163 (1991)

[40] Kumar, A., Zhou, A., Tucker, G., Levine, S.: Conservative q-learning for offline reinforcement learning. Advances in neural information processing systems **33**, 1179–1191 (2020)

[41] Fujimoto, S., Meger, D., Precup, D.: Off-policy deep reinforcement learning without exploration. In: International Conference on Machine Learning, pp. 2052–2062 (2019). PMLR

[42] Chua, K., Calandra, R., McAllister, R., Levine, S.: Deep reinforcement learning in a handful of trials using probabilistic dynamics models. Advances in neural information processing systems **31** (2018)

[43] Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)

[44] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T.P., Silver, D.: Mastering atari, go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020)

[45] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)

[46] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., *et al.*: A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science **362**(6419), 1140–1144 (2018)

[47] Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)

[48] Bellman, R.: Dynamic Programming. Princeton University Press, Princeton, NJ, USA (1957)

[49] Sutton, R.S.: Learning to predict by the methods of temporal differences. Machine Learning **3**(1), 9–44 (1988)

[50] Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning **8**(3–4), 279–292 (1992)

[51] Rummery, G.A., Niranjan, M.: On-line q-learning using connectionist systems. Technical report, University of Cambridge, Department of Engineering (1994)

[52] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: Machine Learning, vol. 8, pp. 229–256. Springer, ??? (1992)

[53] Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)

[54] Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., Freitas, N.: Dueling network architectures for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1995–2003 (2016). PMLR

[55] Schaul, T., Quan, J., Antonoglou, I., Silver, D.: Prioritized experience replay. In: International Conference on Learning Representations (ICLR) (2016)

[56] Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., Silver, D.: Rainbow: Combining improvements in deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[57] Mnih, V., Badia, A.P., Mirza, M., Graves, A., Harley, T., Lillicrap, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937 (2016). PMLR

[58] Schulman, J., Levine, S., Abbeel, P., Jordan, M.I., Moritz, P.: Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1889–1897 (2015)

[59] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[60] Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937 (2016). PmLR

[61] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. In: arXiv Preprint arXiv:1707.06347 (2017). https://arxiv.org/abs/1707.06347

[62] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., *et al.*: Mastering the game of go with deep neural networks and tree search. nature **529**(7587), 484–489 (2016)

[63] Levine, S., Finn, C., Darrell, T., Abbeel, P.: End-to-end training of deep visuomotor policies. Journal of Machine Learning Research **17**, 1334–1373 (2016)

[64] Wu, Y., Tucker, G., Nachum, O.: Behavior regularized offline reinforcement learning. arXiv preprint arXiv:1911.11361 (2019)

[65] Fujimoto, S., Gu, S.S.: A minimalist approach to offline reinforcement learning. Advances in neural information processing systems **34**, 20132–20145 (2021)

[66] Kostrikov, I., Nair, A., Levine, S.: Offline reinforcement learning with implicit q-learning. arXiv preprint arXiv:2110.06169 (2021)

[67] Canese, L., Cardarilli, G.C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., Spanò, S.: Multi-agent reinforcement learning: A review of challenges and applications. Applied Sciences

11(11), 4948 (2021)

[68] Zhang, K., Yang, Z., Başar, T.: Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of reinforcement learning and control, 321–384 (2021)

[69] Buşoniu, L., Babuška, R., De Schutter, B.: Multi-agent reinforcement learning: An overview. Innovations in multi-agent systems and applications-1, 183–221 (2010)

[70] Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S.: Counterfactual multi-agent policy gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[71] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. In: Advances in Neural Information Processing Systems, vol. 30, pp. 6379–6390 (2017)

[72] Foerster, J.N., Chen, R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., Mordatch, I.: Learning with opponent-learning awareness. arXiv preprint arXiv:1709.04326 (2017)

[73] Chow, Y., Tamar, A., Mannor, S., Pavone, M.: Risk-sensitive and robust decision-making: a cvar optimization approach. Advances in neural information processing systems **28** (2015)

[74] Bellemare, M.G., Dabney, W., Munos, R.: A distributional perspective on reinforcement learning. In: International Conference on Machine Learning, pp. 449–458 (2017). PMLR

[75] Dabney, W., Rowland, M., Bellemare, M., Munos, R.: Distributional reinforcement learning with quantile regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[76] Pinto, L., Davidson, J., Sukthankar, R., Gupta, A.: Robust adversarial reinforcement learning. In: International Conference on Machine Learning, pp. 2817–2826 (2017). PMLR

[77] Vinitsky, E., Du, Y., Parvate, K., Jang, K., Abbeel, P., Bayen, A.: Robust reinforcement learning using adversarial populations. arXiv preprint arXiv:2008.01825 (2020)

[78] Kamalaruban, P., Huang, Y.-T., Hsieh, Y.-P., Rolland, P., Shi, C., Cevher, V.: Robust reinforcement learning via adversarial training with langevin dynamics. Advances in Neural Information Processing Systems **33**, 8127–8138 (2020)

[79] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 23–30 (2017). IEEE

[80] Achiam, J., Held, D., Tamar, A., Abbeel, P.: Constrained policy optimization. In: International Conference on Machine Learning, pp. 22–31 (2017). PMLR

[81] Garcıa, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research **16**(1), 1437–1480 (2015)

[82] Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., Topcu, U.: Safe reinforcement learning via shielding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[83] ElSayed-Aly, I., Bharadwaj, S., Amato, C., Ehlers, R., Topcu, U., Feng, L.: Safe multi-agent reinforcement learning via shielding. arXiv preprint arXiv:2101.11196 (2021)

[84] Zhao, R., Li, Y., Fan, Y., Gao, F., Tsukada, M., Gao, Z.: A survey on recent advancements in autonomous driving using deep reinforcement learning: Applications, challenges, and solutions. IEEE Transactions on Intelligent Transportation Systems (2024)

[85] Berkenkamp, F., Turchetta, M., Schoellig, A., Krause, A.: Safe model-based reinforcement learning with stability guarantees. Advances in neural information processing systems **30** (2017)

[86] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)

[87] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)

[88] Hafner, D., Lillicrap, T.P., Norouzi, M., Ba, J.: Mastering atari with discrete world models. In: Conference on Neural Information Processing Systems (2020)

[89] Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.P.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)

[90] Wang, Z., Hunt, J.J., Zhou, M.: Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv preprint arXiv:2208.06193 (2022)

[91] Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., Madry, A.: Implementation matters in deep rl: A case study on ppo and trpo. In: International Conference on Learning Representations (2019)

[92] Granter, S.R., Beck, A.H., Papke Jr, D.J.: Alphago, deep learning, and the future of the human microscopist. Archives of pathology & laboratory medicine **141**(5), 619–621 (2017)

[93] Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)

[94] Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems **36**, 66923–66939 (2023)

[95] Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)

[96] Miao, Z., Wang, J., Wang, Z., Yang, Z., Wang, L., Qiu, Q., Liu, Z.: Training diffusion models towards diverse image generation with reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10844–10853 (2024)

[97] Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8228–8238 (2024)

[98] Deng, F., Wang, Q., Wei, W., Hou, T., Grundmann, M.: Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7423–7433 (2024)

[99] Yang, S., Chen, T., Zhou, M.: A dense reward view on aligning text-to-image diffusion with preference. arXiv preprint arXiv:2402.08265 (2024)

[100] Wu, Z., Gao, H., Wang, Y., Zhang, X., Wang, S.: Universal prompt optimizer for safe text-to-image generation. arXiv preprint arXiv:2402.10882 (2024)

[101] An, J., Zhu, Y., Li, Z., Zhou, E., Feng, H., Huang, X., Chen, B., Shi, Y., Pan, C.: Agfsync: Leveraging ai-generated feedback for preference optimization in text-to-image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 1746–1754 (2025)

[102] Mo, W., Zhang, T., Bai, Y., Su, B., Wen, J.-R., Yang, Q.: Dynamic prompt optimizing for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26627–26636 (2024)

[103] Croitoru, F.-A., Hondru, V., Ionescu, R.T., Sebe, N., Shah, M.: Curriculum direct preference optimization for diffusion and consistency models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 2824–2834 (2025)

[104] Na, S., Kim, Y., Lee, H.: Boost your human image generation model via direct preference optimization. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 23551–23562 (2025)

[105] Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Li, J., Zheng, L.: Step-aware preference optimization: Aligning preference with denoising performance at each step. arXiv preprint arXiv:2406.04314 **2**(5), 7 (2024)

[106] Eyring, L., Karthik, S., Roth, K., Dosovitskiy, A., Akata, Z.: Reno: Enhancing one-step text-to-image models through reward-based noise optimization. Advances in Neural Information Processing Systems **37**, 125487–125519 (2024)

[107] Park, Y.-H., Yun, S., Kim, J.-H., Kim, J., Jang, G., Jeong, Y., Jo, J., Lee, G.: Direct unlearning optimization for robust and safe text-to-image models. Advances in Neural Information Processing Systems **37**, 80244–80267 (2024)

[108] Karthik, S., Coskun, H., Akata, Z., Tulyakov, S., Ren, J., Kag, A.: Scalable ranked preference optimization for text-to-image generation. arXiv preprint arXiv:2410.18013 (2024)

[109] Huang, Q., Chan, L., Liu, J., He, W., Jiang, H., Song, M., Song, J.: Patchdpo: Patch-level dpo for finetuning-free personalized image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 18369–18378 (2025)

[110] Dang, M., Singh, A., Zhou, L., Ermon, S., Song, J.: Personalized preference fine-tuning of diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8020–8030 (2025)

[111] Guo, Z., Zhang, R., Tong, C., Zhao, Z., Gao, P., Li, H., Heng, P.-A.: Can we generate images with cot? let's verify and reinforce image generation step by step. arXiv preprint arXiv:2501.13926 (2025)

[112] Lee, K., Li, X., Wang, Q., He, J., Ke, J., Yang, M.-H., Essa, I., Shin, J., Yang, F., Li, Y.: Calibrated multi-preference optimization for aligning diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 18465–18475 (2025)

[113] Wang, Z., Bao, J., Gu, S., Chen, D., Zhou, W., Li, H.: Designdiffusion: High-quality text-to-design image generation with diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 20906–20915 (2025)

[114] Wu, X., Bai, Y., Zheng, H., Chen, H.H., Liu, Y., Wang, Z., Ma, X., Shu, W.-J., Wu, X., Yang, H., et al.: Lightgen: Efficient image generation through knowledge distillation and direct preference optimization. arXiv preprint arXiv:2503.08619 (2025)

[115] Wang, J., Tian, Z., Wang, X., Zhang, X., Huang, W., Wu, Z., Jiang, Y.-G.: Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. arXiv preprint arXiv:2504.11455 (2025)

[116] Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)

[117] Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)

[118] Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.-A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025)

[119] Zhang, Y., Li, Y., Yang, Y., Wang, R., Yang, Y., Qi, D., Bao, J., Chen, D., Luo, C., Qiu, L.: Reasongen-r1: Cot for autoregressive image generation models through sft and rl. arXiv preprint arXiv:2505.24875 (2025)

[120] Gao, C., Zhang, H., Xu, Z., Cai, Z., Shao, L.: Flip: Flow-centric generative planning as general-purpose manipulation world model. arXiv preprint arXiv:2412.08261 (2024)

[121] Soni, A., Venkataraman, S., Chandra, A., Fischmeister, S., Liang, P., Dai, B., Yang, S.: Videoagent: Self-improving video generation for embodied planning. In: Workshop on Reinforcement Learning Beyond Rewards@ Reinforcement Learning Conference 2025

[122] Ma, Y., Klabjan, D., Utke, J.: Video to video generative adversarial network for few-shot learning based on policy gradient. arXiv preprint arXiv:2410.20657 (2024)

[123] Wu, S., Zhu, Z., Hou, J., Shi, G., Wu, J.: E-motion: Future motion simulation via event sequence diffusion. Advances in Neural Information Processing Systems **37**, 105552–105582 (2024)

[124] Hu, P., Xiao, N., Li, F., Chen, Y., Huang, R.: A reinforcement learning-based automatic video editing method using pre-trained vision-language model. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 6441–6450 (2023)

[125] Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8009–8019 (2025)

[126] Jiang, L., Wu, B., Zhang, J., Guan, X., Chen, S.: Huvidpo: Enhancing video generation through direct preference optimization for human-centric alignment. arXiv preprint arXiv:2502.01690 (2025)

[127] Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)

[128] Wang, S., Tang, H., Dou, Z., Xiong, C.: Harness local rewards for global benefits: Effective text-to-video generation alignment with patch-level reward models. arXiv preprint arXiv:2502.06812 (2025)

[129] Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: Densedpo: Fine-grained temporal preference optimization for video diffusion models. arXiv preprint arXiv:2506.03517 (2025)

[130] Zhang, J., Wu, J., Chen, W., Ji, Y., Xiao, X., Huang, W., Han, K.: Onlinevpo: Align video diffusion model with online video-centric preference optimization. arXiv preprint arXiv:2412.15159 (2024)

[131] Zhang, D., Lan, G., Han, D.-J., Yao, W., Pan, X., Zhang, H., Li, M., Chen, P., Dong, Y., Brinton, C., et al.: Seppo: Semi-policy preference optimization for diffusion alignment. arXiv preprint arXiv:2410.05255 (2024)

[132] Meng, R., Wang, Y., Wu, W., Zheng, R., Li, Y., Ma, C.: Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. arXiv preprint arXiv:2507.03905 (2025)

[133] Cheng, J., Lyu, R., Gu, X., Liu, X., Xu, J., Lu, Y., Teng, J., Yang, Z., Dong, Y., Tang, J., et al.: Vpo: Aligning text-to-video generation models with prompt optimization. arXiv preprint arXiv:2503.20491 (2025)

[134] Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)

[135] Zhu, B., Jiang, Y., Xu, B., Yang, S., Yin, M., Wu, Y., Sun, H., Wu, Z.: Aligning anime video generation with human feedback. arXiv preprint arXiv:2504.10044 (2025)

[136] Furuta, H., Zen, H., Schuurmans, D., Faust, A., Matsuo, Y., Liang, P., Yang, S.: Improving dynamic object interactions in text-to-video generation with ai feedback. arXiv preprint arXiv:2412.02617 (2024)

[137] Yang, X., Tan, Z., Li, H.: Ipo: Iterative preference optimization for text-to-video generation. arXiv preprint arXiv:2502.02088 (2025)

[138] Wang, F.-Y., Shui, Y., Piao, J., Sun, K., Li, H.: Diffusion-npo: Negative preference optimization for better preference aligned generation of diffusion models. arXiv preprint arXiv:2505.11245 (2025)

[139] Yuan, H., Zhang, S., Wang, X., Wei, Y., Feng, T., Pan, Y., Zhang, Y., Liu, Z., Albanie, S., Ni, D.: Instructvideo: Instructing video diffusion models with human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6463–6474 (2024)

[140] Prabhudesai, M., Mendonca, R., Qin, Z., Fragkiadaki, K., Pathak, D.: Video diffusion alignment via reward gradients. arXiv preprint arXiv:2407.08737 (2024)

[141] Kim, J., Kim, B.S., Ye, J.C.: Free^{2}guide: Gradient-free path integral control for enhancing text-to-video generation with large vision-language models. arXiv preprint arXiv:2411.17041 (2024)

[142] Lin, W., Jia, L., Hu, W., Pan, K., Yue, Z., Zhao, W., Chen, J., Wu, F., Zhang, H.: Reasoning physical video generation with diffusion timestep tokens via reinforcement learning. arXiv preprint arXiv:2504.15932 (2025)

[143] Wu, J., Yin, S., Feng, N., Long, M.: Rlvr-world: Training world models with reinforcement learning. arXiv preprint arXiv:2505.13934 (2025)

[144] Fang, X., Ma, L., Chen, Z., Zhou, M., Qi, G.-j.: Inflvg: Reinforce inference-time consistent long video generation with grpo. arXiv preprint arXiv:2505.17574 (2025)

[145] Akizuki, Y., Bernhard, M., Kladeftira, M., Kakooee, R., Dillenburger, B.: Generative modelling with design constraints–reinforcement learning for object generation. In: RE: Anthropocene, Design in the Age of Humans–Proceedings of the 25th CAADRIA Conference, vol. 1, pp. 445–454 (2020). Association for Computer Aided Architectural Design Research in Asia

[146] Zhao, R., Ye, J., Wang, Z., Liu, G., Chen, Y., Wang, Y., Zhu, J.: Deepmesh: Auto-regressive artist-mesh creation with reinforcement learning. arXiv preprint arXiv:2503.15265 (2025)

[147] Liu, J., Xu, J., Guo, S., Li, J., Guo, J., Yu, J., Weng, H., Lei, B., Yang, X., Chen, Z., et al.: Mesh-rft: Enhancing mesh generation via fine-grained reinforcement fine-tuning. arXiv preprint arXiv:2505.16761 (2025)

[148] Zhang, Z., Han, X., Dong, B., Li, T., Yin, B., Yang, X.: Point cloud scene completion with joint color and semantic estimation from single rgb-d image. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 11079–11095 (2023)

[149] Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)

[150] Zhou, Z., Xia, X., Ma, F., Fan, H., Yang, Y., Chua, T.-S.: Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. arXiv preprint

arXiv:2502.04370 (2025)

[151] Zou, X., Xia, R., Wang, H., Zhou, P.: Dreamcs: Geometry-aware text-to-3d generation with unpaired 3d reward supervision. arXiv preprint arXiv:2506.09814 (2025)

[152] Dong, Y., Zuo, Q., Gu, X., Yuan, W., Zhao, Z., Dong, Z., Bo, L., Huang, Q.: Gpld3d: Latent diffusion of 3d shape generative models by enforcing geometric and physical priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 56–66 (2024)

[153] Liu, Q., Liu, Z., Zhang, D., Jia, K.: Nabla-r2d3: Effective and efficient 3d diffusion alignment with 2d rewards. arXiv preprint arXiv:2506.15684 (2025)

[154] Xie, D., Li, J., Tan, H., Sun, X., Shu, Z., Zhou, Y., Bi, S., Pirk, S., Kaufman, A.E.: Carve3d: Improving multi-view reconstruction consistency for diffusion models with rl finetuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6369–6379 (2024)

[155] Choi, J.S., Lee, K., Lee, D., Shin, J., Lee, K.: Hfdream: Improving 3d generation via human-assisted multi-view text-to-image models (2024)

[156] Wang, W., Xu, H., Yang, Y., Liu, Z., Meng, J., Wang, H.: Mvreward: Better aligning and evaluating multi-view diffusion models with human preferences. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 7898–7906 (2025)

[157] Zhang, Z., Shen, L., Ye, D., Luo, Y., Zhao, H., Zhang, L.: Refining few-step text-to-multiview diffusion via reinforcement learning. arXiv preprint arXiv:2505.20107 (2025)

[158] Zhang, R., Gao, W., Li, G., Li, T.H.: Qinet: Decision surface learning and adversarial enhancement for quasi-immune completion of diverse corrupted point clouds. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)

[159] Ostonov, A., Wonka, P., Michels, D.L.: Rlss: A deep reinforcement learning algorithm for sequential scene generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2219–2228 (2022)

[160] Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14738–14749 (2023)

[161] Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11050–11059 (2022)

[162] Garg, A., Zeng, L., Tsarov, A., Kalantari, N.K.: Text2stereo: Repurposing stable diffusion for stereo generation with consistency rewards. arXiv preprint arXiv:2506.05367 (2025)

8

声明
资助
本研究没有收到外部资助。

利益冲突/竞争利益
作者声明不存在竞争利益。
数据可用性：不适用。

代码可用性
不适用。

作者贡献

量子负责人编写了调查结构，规划了所有部分内容，并领导了写作过程。Liang 撰写了摘要、介绍、强化学习理论部分以及关于机制、见解和未来工作的总结部分。他还润色了整个手稿。Yijie Fang 撰写了关于图像生成强化学习的部分。Rui Li 和 Ziqi Ni 共同撰写了视频生成强化学习部分。Ruijie Su 负责 3D/4D 内容生成强化学习的部分。Chi Zhang 和 Xuelong Li 提供了项目监督和整体指导。

感谢作者要感谢同事们在准备这篇综述期间所提供的宝贵反馈。