世界建模的扩散：Atari[†]中的视觉细节问题

埃洛伊-阿隆索[*]
日内瓦大学

亚当-杰利[*]
爱丁堡大学

文森特-米切利
日内瓦大学

安西-卡内维斯托
微软研究院

阿莫斯-斯托基
爱丁堡大学

蒂姆-皮尔斯[‡]
微软研究院

François Fleuret[‡]
日内瓦大学

摘要

世界模型是一种很有前途的方法，可用于以安全、样本效率高的方式训练强化学习代理。最新的世界模型主要通过离散潜变量序列来模拟环境动态。然而，这种压缩为紧凑的离散表示法可能会忽略对强化学习非常重要的视觉细节。与此同时，扩散模型已成为图像生成的主流方法，对离散潜变量建模的成熟方法提出了挑战。在这一模式转变的推动下，我们推出了 DIAMOND（DIffusion As a Model Of eNvironment Dreams），这是一种在扩散世界模型中训练的强化学习代理。我们分析了使扩散适合世界建模所需的关键设计选择，并演示了改进视觉细节如何提高代理性能。在竞争激烈的 Atari 100k 基准测试中，DIAMOND 获得了 1.46 的人类标准化平均分；这是完全在世界模型中训练的代理的新最佳成绩。通过在静态的*《反恐精英：全球攻势》*游戏中进行训练，我们进一步证明了 DIAMOND 的扩散世界模型可以独立成为一个交互式神经游戏引擎*：全球攻势*》游戏中进行训练。为了促进未来对扩散世界建模的研究，我们在 https://diamond-wm.github.io 上发布了我们的代码、代理、视频和可播放的世界模型。

1 导言

环境生成模型或 "世界模型"（Ha 和 Schmidhuber，2018 年）作为通用代理规划和推理其环境的组成部分，正变得越来越重要（LeCun，2022 年）。近年来，强化学习（RL）取得了广泛的成功（Silver 等人，2016；Degrave 等人，2022；欧阳等人，2022），但众所周知，强化学习的样本效率较低，这限制了其在现实世界中的应用。世界模型已显示出在不同环境中训练强化学习代理的前景（Hafner 等人，2023 年；Schrittwieser 等人，2020 年），并大大提高了样本效率（Ye 等人，2021 年），从而可以在现实世界中从经验中学习（Wu 等人，2023 年）。

最近的世界建模方法（Hafner 等人，2021 年；Micheli 等人，2023 年；Robine 等人，2023 年；Hafner 等人，2023 年；Zhang 等人，2023 年）通常将环境动态建模为离散潜变量序列。潜变量空间的离散化有助于避免多步时间跨度上的复合误差。然而，这种编码可能会丢失信息，导致通用性和重建质量下降。这在现实世界中可能会出现问题，因为在这种情况下，信息

(为避免混淆，本文是（阿隆索等，2023 年）的最终版本，与（丁等，2024 年）无关。)

[*]等效贡献。[‡]平等监督。联系方式：eloi.alonso@unige.ch 和 adam.jelley@ed.ac.uk

第 38 届神经信息处理系统大会（NeurIPS 2024）。

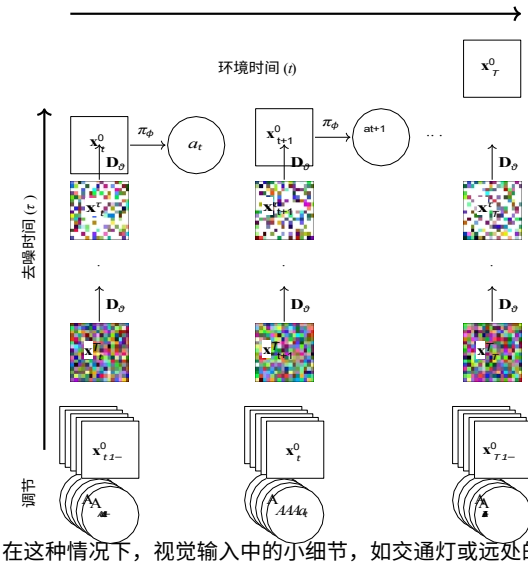


图 1：DIAMOND 随时间展开的想象。上一行描述了策略 π_j在我们学习的扩散世界模型 $\mathbf{D}_\vartheta$的想象中采取的一系列行动。环境时间 t 沿着横轴流动，而纵轴代表从 T 到 0 的去噪时间 τ。

$\mathbf{x}^0$、行动 $a_{\leqslant t}$，并从
在初始噪声样本 $\mathbf{x}^T$的基础上，我们通过反复调用
噪声过程 $\{\mathbf{x}^{(\tau)}\}(\tau)$ $(=0)$
$\mathbf{D}$
$_\vartheta$，得到（干净的）下一个观测值 $\mathbf{x}_t$。
想象过程是自回归的
即预测的观测值 $\mathbf{x}^0$和政策采取的行动 a_t成为下一时间步的条件的一部分。有关这一过程的可视化动画，请访问 https://diamond-wm.github.io。

在这种情况下，视觉输入中的小细节，如交通灯或远处的行人，可能会改变代理的策略。在这种情况下，视觉输入中的小细节，如远处的红绿灯或行人，可能会改变代理的策略。增加离散潜变量的数量可以减轻这种有损压缩，但同时也会增加计算成本（Micheli 等人，2023 年）。

与此同时，扩散模型（Sohl-Dickstein 等人，2015 年；Ho 等人，2020 年；Song 等人，2020 年）已成为高分辨率图像生成的主流范式（Rombach 等人，2022 年；Podell 等人，2023 年）。这一类方法中，模型学会了逆向噪声过程，对离散标记建模的成熟方法提出了挑战（Esser 等人，2021 年；Ramesh 等人，2021 年；Chang 等人，2023 年），从而为减轻世界建模中的离散化需求提供了一种有前途的替代方法。此外，众所周知，扩散模型易于条件化，可以灵活地模拟复杂的多模式分布，而不会出现模式崩溃。这些特性对世界建模非常重要，因为遵守条件应该能让世界模型更贴近地反映代理的行为，从而使学分分配更可靠，而多模态分布建模应该能为代理提供更多样化的训练场景。

受这些特点的启发，我们提出了 DIAMOND（DIffusion As a Model Of eNvironment Dreams），一个在扩散世界模型中训练的强化学习代理。为了确保我们的扩散世界模型在长时间内高效稳定，我们必须做出谨慎的设计选择，我们将通过定性分析来说明这些选择的重要性。在久负盛名的 Atari 100k 基准测试中，DIAMOND 获得了 1.46 的人类平均归一化分数；这对于完全在世界模型中训练的代理来说是一个全新的技术水平。此外，在图像空间中运行的好处还在于，我们的扩散世界模型可以直接替代环境，从而为世界模型和代理行为提供更深入的了解。特别是，我们发现在某些游戏中，由于对关键视觉细节进行了更好的建模，性能得到了提高。为了进一步证明世界模型的有效性，我们在 87 小时的静态*《反恐精英：全球攻势（CSGO）》*游戏中训练了 DIAMOND 的扩散世界模型*：全球攻势*》（CSGO）游戏（Pearce 和 Zhu，2022 年）中训练 *DIAMOND* 的扩散世界模型，为流行的游戏地图《*尘埃 II*》制作一个交互式神经游戏引擎。我们在 `https://diamond-wm.github.io` 上发布了我们的代码、代理和可玩世界模型。

2 前言

2.1 强化学习和世界模型

我们将环境建模为标准的部分可观测马尔可夫决策过程（POMDP）（Sutton and Barto, 2018），即（S, A, O, T, R, O, γ），其中 S 是一组状态，A 是一组离散行动，O 是一组图像观测。过渡函数 T : S ×A×S→ [0, 1] 描述了环境动态 $p(\mathbf{s}_{(t)\,(+1)}|\ \mathbf{s}_{(t)}, \mathbf{a}_{(t)})$，奖励函数 R : S× A× S → R 将过渡映射到标量奖励。代理不能直接访问状态 s_t，只能通过图像观测 x_t∈ O 看到环境，而图像观测是根据观测概率 $p(\mathbf{x}_t|\ \mathbf{s}_{(t)}$发布的、）

由观测函数 $O : S \times O \to [0, 1]$ 描述。目标是得到一个政策π，将观察结果映射到行动，以最大化预期贴现收益 $E_{\pi}[\sum_{t\geq 0} \gamma^t r_t]$，其中 $\gamma \in [0, 1]$ 是一个贴现因子。世界模型（Ha 和 Schmidhuber，2018）是环境的生成模型，即 $p(s_{t+1}, r_t | s_t, a_t)$的模型。）这些模型可用作模拟环境，以样本效率高的方式训练 RL 代理（Sutton，1991 年）（Wu 等人，2023 年）。在这种模式下，训练过程通常包括以下三个循环步骤：在真实环境中使用 RL 代理收集数据；在所有收集到的数据上训练世界模型；在世界模型环境中训练 RL 代理（通常称为 "想象中"）。

2.2 基于分数的扩散模型

扩散模型（Sohl-Dickstein 等人，2015 年）是一类受非平衡态热力学启发的生成模型，它通过逆转噪声过程生成样本。

我们考虑一个扩散过程 $\{\mathbf{x}^\tau\}_{\tau \in [0,T]}$，以连续时间变量 $\tau \in [0,T]$为索引，相应的边际值$\{p^\tau\}_{\tau \in [0,T]}$ 以及边界条件 $p(0)$ $p(\text{data})$ 和 $p(T)$ $p^{(prior)}$，其中 $p^{(prior)}$是可控的非结构化先验分布，如 $\mathrm{Ga}(0, 1)$，以及边界条件 $p^0 = p^{数据}$和 $p^T = p^{(先验)}$，其中 $p^{先验}$是一个可行的非结构化先验分布，如高斯分布。请注意，我们使用 τ 和上标表示扩散过程时间，以便保留 t 和下标表示环境时间。

这一扩散过程可以描述为标准随机微分方程（SDE）的解（Song 等人，2020 年）、

$$\mathbf{dx} = \mathbf{f}(\mathbf{x}, \tau)d\tau + g(\tau)\mathbf{dw}, \tag{1}$$

其中，$\mathbf{w}$ 是维纳过程（布朗运动），$\mathbf{f}$ 是作为漂移系数的矢量值函数，g 是称为过程扩散系数的标量值函数。

要获得从噪声映射到数据的生成模型，我们必须逆转这一过程。值得注意的是，安德森（Anderson，1982 年）指出，反向过程也是一个扩散过程，在时间上向后运行，用下面的 SDE 描述、

$$\mathbf{dx} = [\mathbf{f}(\mathbf{x}, \tau) - g(\tau)^2 \nabla_{\mathbf{x}} \log p^{(\tau)}(\mathbf{x})]d\tau + g(\tau)\bar{\mathbf{dw}}, \tag{2}$$

其中，$\bar{\mathbf{w}}$是反向时间维纳过程，$\nabla_{\mathbf{x}} \log p^\tau(\mathbf{x})$ 是（斯坦因）得分函数，即对数边际值相对于支持度的梯度。因此，为了反转前向噪声过程，我们只需定义函数 f 和 g（见第 3.1 ），节并估算与沿过程的边际值 $\{p^{(\tau)}\}_{\tau \in [0,T]}$ 相关的未知得分函数$\nabla_{\mathbf{x}} \log p^\tau(\mathbf{x})$。实际上，可以使用单一的随时间变化的分数模型 $\mathbf{S}_\vartheta(\mathbf{x}, \tau)$ 来估计这些分数函数（Song 等人，2020 年）。

由于我们无法获得真实的分数函数，因此在任何时候估计分数函数都不是一件容易的事。幸运的是，Hyvärinen（2005 年）引入了*分数匹配*目标，该目标竟然可以在不知道底层分数函数的情况下从数据样本中训练分数模型。要从边际 p^τ中获取样本，我们需要模拟从时间 0 到时间 τ 的前向过程，因为我们只有干净的数据样本。一般来说，这样做的成本很高，但如果 f 是仿射的，我们就可以通过对 干净数据样本应用高斯扰动核 $p^{0\tau}$*来*分析前向过程中的任意时间 τ，只需一步就能达到（Song 等人，2020 年）。由于核是可变的，因此分数匹配简化为*去噪分数匹配*目标（Vincent，2011）、

$$L(\theta) = E\left\| \mathbf{S}_\vartheta(\mathbf{x}^\tau, \tau) - \nabla_{\mathbf{x}^\tau} \log p^{0\tau}(\mathbf{x}^\tau | \mathbf{x}^0) \right\|^2, \tag{3}$$

其中，期望值是在扩散时间 τ 内，通过对干净样本 $\mathbf{x}^0 \sim p^{data}(\mathbf{x}^0)$ 应用 τ 级扰动核得到的噪声样本 $\mathbf{x}^\tau \sim p^{0\tau}(\mathbf{x}^\tau | \mathbf{x}^0)$。重要的是，由于核 $p^{0\tau}$是已知的高斯分布，这一目标变成了简单的 L_2重建损失、

$$L(\theta) = E\left\| \mathbf{D}_\vartheta(\mathbf{x}^\tau, \tau) - \mathbf{x}^0 \right\|^2, \tag{4}$$

重新参数化 $\mathbf{D}_\vartheta(\mathbf{x}^\tau, \tau) = \mathbf{S}_\vartheta(\mathbf{x}^\tau, \tau)\sigma^2(\tau) + \mathbf{x}^\tau$，其中 $\sigma(\tau)$是 τ 级扰动核的方差。

2.3 世界模型的扩散

第 2.2 节中描述的基于分数的扩散模型提供了 $p_{数据}$的无条件生成模型。为了作为世界模型，我们需要一个环境动态的条件生成模型，即 $p(\mathbf{x}_{t+1}|\ \mathbf{x}_{\leq t},\ a_{\leq t})$，这里我们考虑 POMDP 的一般情况，其中马尔可夫状态 s_t是未知的，可以通过过去的观测和行动来近似。如图 1 。所示，我们可以在此历史上建立一个扩散模型，直接估计并生成下一个观测值这对方程 4 修改如下

$$\mathsf{L}(\theta) = \mathsf{E}\left\| \mathbf{D}_\vartheta(\mathbf{x}^\tau \quad , \tau, \mathbf{x}^0_{t+1},\ a_{\leq t}) - \mathbf{x}^{(0)}_{\quad t+1} \right\|_2. \tag{5}$$

在训练过程中，我们从回放数据集中采样一个轨迹段 $\mathbf{x}^0$、$\overline{\leq t}$个$_{\leq t}$、$\mathbf{x}^{(0)}_{t+1}$ 和
我们就能得到下一个观测值 $\mathbf{x}^\tau$ $\sim p^{(0)\,(\tau)\,(x\tau}$ $|\mathbf{x}^{\ 0}_{t+1})$，应用 τ 级扰动
核。总之，世界建模的扩散过程类似于第 2.2 节中描述的标准扩散过程，其分数模型以过去的观察和行动为条件
。

如图 1 。所示，为了对下一个观察结果进行采样，我们需要迭代求解方程 2 中的反向 SDE虽然原则上我们可以使用任何 ODE 或 SDE 求解器，但在采样质量和函数求值次数（NFE）之间存在固有的权衡，这直接决定了扩散世界模型的推理成本（详见附录 A）。

3 方法

3.1 扩散范式的实际选择

在第 2 ，节提供的背景基础上我们现在介绍 DIAMOND，将其作为基于扩散的世界模型的 实际实现。特别是，我们现在定义第 2.2 节中介绍的漂移和扩散系数 $\mathbf{f}$ 和 g，它们与特定的扩散范式选择相对应。虽然 DDPM（Ho 等人，2020 年）是此类选择的一个例子（如附录 B 所述），而且历来是自然的候选范例，但我们采用了 Karras 等人（2022 年）提出的 EDM 模式。这一选择的实际影响将在第 5.1 节中讨论。

5.1.接下来，我们将介绍如何调整 EDM 以建立基于扩散的世界模型。

我们考虑扰动核 $p^{(0)\,(\tau)\,(x\tau}$ $_{t+1}|\mathbf{x}^{\ 0}_{t+1})= \mathsf{N}(\mathbf{x}^\tau$ $_{t+1}; \mathbf{x}^{\ 0}_{t+1}, \sigma^{(2)\,(\tau)}\mathbf{I})$，其中 $\sigma(\tau)$ 是一个实值函数，称为噪声时间表。
是扩散时间的实值函数，称为噪声时间表。 这相当于将漂移和扩散系数设为 $\mathbf{f}(\mathbf{x}, \tau) = \mathbf{0}$（仿射）和 $g(\tau) = 2\dot{\sigma}(\tau)\sigma(\tau)$。

我们使用 Karras 等人（2022 年）引入的网络预处理，因此将公式 5 中的 $\mathbf{D}_\vartheta$参数化为噪声观测值与神经网络 $\mathbf{F}_\vartheta$预测值的加权和、

$$\mathbf{D}_{(\vartheta)\,(}\mathbf{x}^\tau_{t+1}\ ,\ y^{(\tau)}) = c^\tau_{跳过}\ \mathbf{x}^\tau_{t+1}\ + c^\tau_{出}\mathbf{F}_\vartheta c^\tau \mathbf{x}^{(\tau)}_{进}\ ,_{t+1}, \tau, y^{(\tau)}、_{\ t} \tag{6}$$

为简洁起见，我们定义 $y^\tau := (c^{\tau)}_{\quad t}$ $_{噪声}$，$\mathbf{x}^0_{\leq t}\ a_{\leq t})$ 包括所有条件变量。
先决条件器 c^τ和 c^τ 的选择是为了使网络的输入和输出在以下情况下保持单位方差
任何噪声水平 $\sigma(\tau), c^\tau$ $_{噪声}$ 是噪音水平的经验变换，而 c^τ $_{跳过}$ 是以
的 $\sigma(\tau)$ 和数据分布的标准偏差 σ_{data}，如 c^τ $_{跳过} = \sigma^{\ 2}_{数据} / (\sigma^{\ 2}_{数据} + \sigma^{(2)\,(\tau)})$.
附录 C 全面介绍了这些前置条件。

结合等式 5 和 6 可以了解 $\mathbf{F}_\vartheta$的训练目标、

$$\mathsf{L}(\theta) = \mathsf{E}\left\| \mathbf{F}_{\underbrace{c^{(\tau)}\mathbf{x}^\tau\ ,\ y^\tau}_{网络p\ \ 重新预测}} - \underbrace{c^{(\tau)\,x(\tau}_{out}\ \ _{t+1} - c^{(\tau)\,x(\tau}_{跳过}\ \ _{t+1}}_{网络训练目标} \right\|_{(2)} \tag{7}$$

网络训练目标根据退化程度自适应地混合信号和噪声
$\sigma(\tau)$.当 $\sigma(\tau) \gg \sigma_{数据}$时，我们有 c^τ $_{跳过}$ →为 0，$\mathbf{F}_\vartheta$的训练目标由干净的
信号 $\mathbf{x}^0$占主导地位。相反，当噪声电平较低时，$\sigma(\tau) \to 0$，则 $c^{(\tau)}$ $_{跳过}$ →为 1，目标
成为干净信号和扰动信号之间的差值，即添加的高斯噪声。直观地说，这可以防止训练目标在低噪声条件下变得

微不足道。在实践中，这一目标在噪声表的极端情况下会出现高方差，因此 Karras 等人（2022 年）从根据经验选择的对数正态分布中对噪声水平 $\sigma(\tau)$ 进行采样，以便将训练集中在中等噪声区域，如附录 C 所述。

我们对向量场 $\mathbf{F}_\theta$使用标准的 U-Net 2D（Ronneberger 等人，2015 年），并保留 L 个过去观察和行动的缓冲区，用于调节模型。我们将这些过去的观测数据与下一个噪声观测数据按通道串联起来，并通过 U-Net 的残差块（He 等人，2015 年）中的自适应组归一化层（Zheng 等人，2020 年）输入动作。

正如第 2.3 节和附录 A 所述，有许多可能的采样方法可以从训练有素的扩散模型中生成下一个观测值。虽然我们的代码库支持多种采样方案，但我们发现欧拉方法非常有效，既不会产生高阶采样器所需的额外 NFE 成本，也不会产生随机采样的不必要复杂性。

3.2 想象中的强化学习

有了第 3.1 ，节中的扩散模型我们现在用奖励和终止模型来完善我们的世界模型，这也是在想象中训练 RL 代理所必需的。由于估计奖励和终止是标量预测问题，我们使用了由标准 CNN（LeCun 等人，1989 年；He 等人，2015 年）和 LSTM（Hochreiter 和 Schmidhuber，1997 年；Gers 等人，2000 年）层组成的单独模型 R_ψ来处理部分可观测性。RL 代理包括一个由共享 CNN-LSTM 参数化的演员-批评网络，该网络具有策略和值头。策略π_j是用 REINFORCE 和价值基线来训练的，我们使用贝尔曼误差（λ-*returns*）来训练价值网络 V_j，类似于 Micheli 等人（2023 年）的做法。如第 2.1 节所述，我们完全在想象中训练代理。代理只在数据收集时与真实环境互动。每个收集阶段结束后，当前的世界模型会通过对迄今为止收集到的所有数据进行训练而得到更新。然后，在更新后的世界模型环境中使用 RL 对代理进行训练，并重复上述步骤。该过程详见算法 1，与 Kaiser 等人（2019 年）、Hafner 等人（2020 年）和 Micheli 等人（2023 年）的算法相似。我们在附录 D、E、F 中分别提供了架构细节、超参数和 RL 目标。

4 实验

4.1 雅达利 100k 基准

为了对 DIAMOND 进行全面评估，我们使用了已建立的 Atari 100k 基准（Kaiser 等人，2019 年），该基准由 26 个游戏组成，可测试各种代理能力。每个游戏只允许代理在环境中进行 100k 次操作，这大致相当于人类 2 小时的游戏时间，以便在评估前学习如何玩游戏。作为参考，无约束 Atari 代理通常要训练 5000 万步，经验增加了 500 倍。我们对 DIAMOND 进行了从头开始的训练，每款游戏有 5 个随机种子。每次运行使用约 12GB 的 VRAM，在单个 Nvidia RTX 4090 上耗时约 2.9 天（总共 1.03 GPU 年）。

在表 1 中，我们将其与完全在世界模型中训练代理的其他最新方法进行了比较，包括 STORM（Zhang 等人，2023 年）、DreamerV3（Hafner 等人，2023 年）、IRIS（Micheli 等人，2023 年）、TWM（Robine 等人，2023 年）和 SimPle（Kaiser 等人，2019 年）。附录 J 提供了与无模型方法和基于搜索方法的更广泛比较，包括 BBF（Schwarzer 等，2023 年）和 EfficientZero（Ye 等，2021 年），它们是目前在该基准上表现最好的方法。BBF 和 EfficientZero 使用的技术是正交的，与我们的方法没有直接可比性，例如 BBF 结合超参数调度使用周期性网络重置，EfficientZero 使用计算昂贵的前瞻蒙特卡洛树搜索。将这些附加组件与我们的世界模型相结合，将是未来工作的一个有趣方向。

4.2 雅达利 100k 基准测试结果

表 1 提供了所有游戏的得分以及人类标准化得分（HNS）的平均值和四分位数平均值（IQM）（Wang 等人，2016 年）。根据 Agarwal 等人（2021 年）关于点估计局限性的建议，我们提供了分层自举置信区间。

平均值和 IQM 值见图 2，性能概况和其他指标见附录 H。

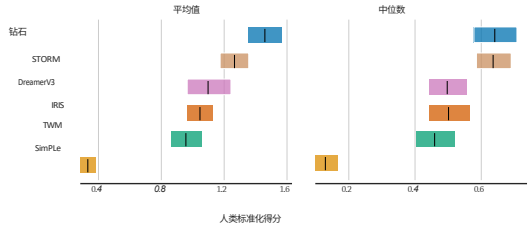


图 2：人类标准化得分的平均值和四分位数之间的平均值。蓝色 DIAMOND 的 HNS 平均值为 1.46，IQM 为 0.64。

表 1：经过 2 小时实时体验后，Atari 100k 基准中 26 款游戏的收益率以及人类标准化综合指标。粗体数字表示性能最好的方法。从 5 个种子的平均得分来看，DIAMOND 明显优于其他世界模型基线。

游戏	随机	人类	模拟	TWM	IRIS	DreamerV3	风暴	钻石（我们的）
异形	227.8	7127.7	616.9	674.6	420.0	959.0	**983.6**	744.1
阿米达尔	5.8	1719.5	74.3	121.8	143.0	139.0	204.8	**225.8**
攻击	222.4	742.0	527.2	682.6	1524.4	706.0	801.0	**1526.4**
亚力	210.0	8503.3	1128.3	1116.6	853.6	932.0	1028.0	**3698.5**
银行抢劫	14.2	753.1	34.2	466.7	53.1	**649.0**	641.2	19.7
战区	2360.0	37187.5	4031.2	5068.0	13074.0	12250.0	**13540.0**	4702.0
拳击	0.1	12.1	7.8	77.5	70.1	78.0	79.7	**86.9**
突破	1.7	30.5	16.4	20.0	83.7	31.0	15.9	**132.5**
斩波器指挥	811.0	7387.8	979.4	1697.4	1565.0	420.0	**1888.0**	1369.8
疯狂攀岩	10780.5	35829.4	62583.6	71820.4	59324.2	97190.0	66776.0	**99167.8**
恶魔攻击	152.1	1971.0	208.1	350.2	**2034.4**	303.0	164.6	288.1
高速公路	0.0	29.6	16.7	24.3	31.1	0.0	**33.5**	33.3
冻伤	65.2	4334.7	236.9	**1475.6**	259.1	909.0	1316.0	274.1
地鼠	257.6	2412.5	596.8	1674.8	2236.1	3730.0	**8239.6**	5897.9
英雄	1027.0	30826.4	2656.6	7254.0	7037.4	**11161.0**	11044.3	5621.8
詹姆斯邦德	29.0	302.8	100.5	362.4	462.7	445.0	**509.0**	427.4
袋鼠	52.0	3035.0	51.2	1240.0	838.2	4098.0	4208.0	**5382.2**
克拉尔	1598.0	2665.5	2204.8	6349.2	6616.4	7782.0	8412.6	**8610.1**
KungFuMaster	258.5	22736.3	14862.5	24554.6	21759.8	21420.0	**26182.0**	18713.6
MsPacman	307.3	6951.6	1480.0	1588.4	999.1	1327.0	**2673.5**	1958.2
庞	-20.7	14.6	12.8	18.8	14.6	18.0	11.3	**20.4**
私家侦探	24.9	69571.3	35.0	86.6	100.0	882.0	**7781.0**	114.3
Qbert	163.9	13455.0	1288.8	3330.8	745.7	3405.0	**4522.5**	4499.3
RoadRunner	11.5	7845.0	5640.6	9109.0	9614.6	15565.0	17564.0	**20673.2**
海洋探索	68.4	42054.7	683.3	**774.4**	661.3	618.0	525.2	551.2
向上向下	533.4	11693.2	3350.3	**15981.7**	3546.2	9234.0	7985.0	3856.3
#Superhuman ( )个	0	不适用	1	8	10	9	10	**11**
平均值 ( )个	0.000	1.000	0.332	0.956	1.046	1.097	1.266	**1.459**
IQM ( )个	0.000	1.000	0.130	0.459	0.501	0.497	**0.636**	**0.641**

我们的研究结果表明，DIAMOND 在所有基准测试中都表现出色，在 11 个对局中表现优于人类棋手，并取得了 1.46 的超人平均 HNS 值，在完全在世界模型中训练的代理中创造了新的最佳成绩。DIAMOND 的 IQM 也与 STORM 不相上下，并且高于所有其他基准。我们发现 DIAMOND 在捕捉小细节非常重要的环境中表现尤为出色，例如《*Asterix* 》、《*Breakout*》和《*Road Runner*》。我们将在第 5.3 。节中进一步对世界模型的视觉质量进行定性分析

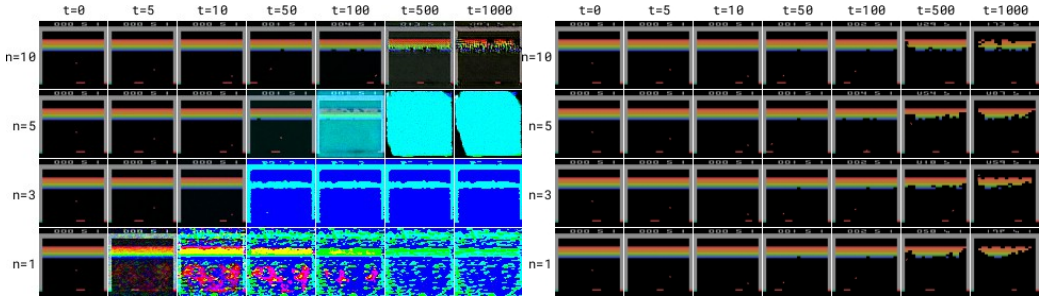
5 分析

5.1 扩散框架的选择

如第 2 ，节所述我们原则上可以在世界模型中使用任何扩散模型变体。DIAMOND 采用了第 3 ，而 节所述的 EDM（Karras 等人，2022 年）DDPM（Ho 等人，2020 年）也是一个自然的候选模型，因为它已被用于许多图像生成应用中（Rombach 等人，2022 年；Nichol 和 Dhariwal，2021 年）。我们将在本节证明这一设计决定的合理性。

为了将 DDPM 与我们的 EDM 实现进行公平比较，我们使用相同的网络架构，在共享的 100k 帧静态数据集上对两种变体进行了训练，这些数据集是在游戏 "*突围*"（*Breakout*）中使用专家策略收集的。正如第 2.3 ，节中所讨论的去噪步骤的数量与世界模型的推理成本直接相关，因此减少去噪步骤将降低根据想象轨迹训练代理的成本。Ho 等人（2020）使用了上千个去噪步骤，Rombach 等人（2022）使用了数百个稳定扩散步骤。然而，为了使我们的世界模型在计算上与其他世界模型基线（如 IRIS，其每个时间步需要 16 个 NFE）相当，我们最多需要几十个去噪步骤，最好更少。遗憾的是，如果去噪步数设置得太少，视觉质量就会下降，从而导致误差加剧。

为了研究扩散变体的稳定性，我们在图 3 中展示了在不同的去噪步数 $n \leq 10$ 下，自回归生成的想象轨迹，直到 $t = 1000$ 个时间步。我们看到，在这种情况下使用 DDPM（图 3a）会导致严重的复合误差，使世界模型迅速偏离分布。相比之下，基于 EDM 的扩散世界模型（图 3b）在较长的时间跨度内显得更加稳定，即使是单一去噪步骤也是如此。附录 K 提供了对这种复合误差的定量分析。



(a) 基于 DDPM 的世界模型轨迹。 (b) 基于 EDM 的世界模型轨迹。

图 3：基于 DDPM（左）和 EDM（右）的扩散世界模型的想象轨迹。= 我们观察到，基于 DDPM 的生成存在复合误差，而且去噪步数越少，误差累积的速度越快。相比之下，我们基于 EDM 的世界模型显得更加稳定，即使在 $n = 1$ 的情况下也是如此。

与 DDPM 采用的更简单的噪声预测目标相比，等式 7 所描述的训练目标经过了改进，从而产生了这一令人惊讶的结果。虽然预测噪声在中等噪声水平下效果很好，但这一目标会导致模型学习标识函数 $\xi_\theta(\mathbf{x}^\tau_{t+1}, y^\tau_t) \to \mathbf{x}^\tau_t$，当噪声占主导地位时（$\sigma_{noise} \gg \sigma_{data} \Rightarrow \xi_\theta(\mathbf{x}^\tau_{t+1}, y^\tau_t) \to \mathbf{x}^\tau_t$），其中 $\xi_\theta(\mathbf{x}^\tau_{t+1}, y^\tau_t)$ 是噪声预测网络。这样，在采样程序开始时，对分数函数的估计就会很差，从而降低生成质量，导致复合误差。

与此相反，第 3.1 节所述的 EDM 所采用的信号和噪声自适应混合方法意味着，当噪声占主导地位时，模型经过训练可预测干净图像（$\sigma_{噪声} \gg \sigma_{data} \Rightarrow \mathbf{F}_\theta(\mathbf{x}^\tau_{t+1}, y^\tau_t) \to \mathbf{x}^0_t$）。这样就能更好地估计在没有 $\mathbf{x}^\tau_{t+1}$ 如图 3b 所示。

5.2 去噪步骤数的选择

如图 3b，但最后一行所示，我们发现基于 EDM 的世界模型在只有一个去*噪*步骤的情况下非常稳定我们在此讨论了这一选择在某些情况下会如何限制模型的视觉质量。我们将在附录 L 中提供更多定量分析。

如第 2.2，节所述我们的得分模型等同于用 L_2 重构损失训练的去噪自编码器（Vincent 等人，2008 年）。因此，最佳单步预测是对给定噪声输入的可能重构的期望，如果该后验分布是多模态的，则可能超出分布范围。有些游戏（如 *"突围"*）的转换是确定的，只需一步去噪就能准确建模（见图 3b），而在其他一些游戏中，部分可观测性会导致多模态观测分布。在这种情况下，就需要使用迭代求解器来驱动采样程序向特定模式移动，如图 4 。中的游戏*方框*所示因此，我们在所有实验中都将 $n = 3$。

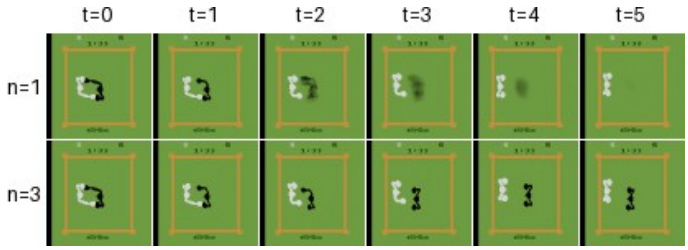
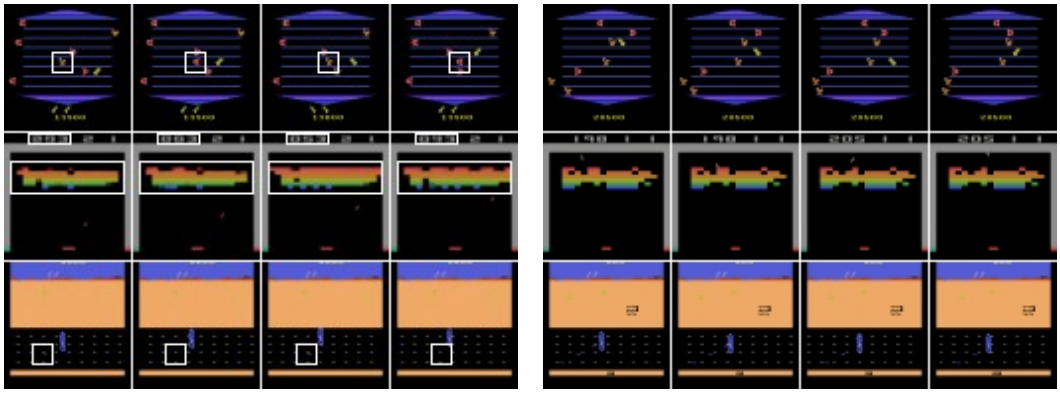


图 4：*拳击*比赛中的单步采样（上行）与多步采样（下行）。黑方选手的移动是不可预测的，因此单步去噪会在可能的结果之间进行插值，导致预测结果模糊不清。与此相反，多步采样通过驱动生成器向特定模式移动，从而生成清晰的图像。有趣的是，政策控制着白方棋手，因此世界模型知道他的行动。这一信息消除了任何模糊性，因此我们观察到单步采样和多步采样都能正确预测白人玩家的位置。

5.3 与 IRIS 的定性视觉比较

我们现在将其与 IRIS（Micheli 等人，2023 年）进行比较，IRIS 是一种成熟的世界模型，它使用离散自动编码器（Van Den Oord 等人，2017 年）将图像转换为离散标记，并使用自回归变换器（Radford 等人，2019 年）将这些标记随时间进行合成。为了进行公平比较，我们在使用专家策略收集的 100k 帧相同静态数据集上对两个世界模型进行了训练。比较结果如下图 5 所示。



(a) IRIS　　(b) 钻石

图 5：IRIS（左）和 DIAMOND（右）所想象的连续帧。白色方框突出显示了帧与帧之间的不一致性，只有在使用 IRIS 生成的轨迹中才会出现这种情况。在《*Asterix*》（上排）中，敌人（橙色）在第二帧变为奖励（红色），然后在第三帧变回敌人，在第四帧再次变回奖励。在 *Breakout*（中排）中，砖块和分数在不同帧之间不一致。在《*Road Runner*》（最下面一行）中，奖励（路上的小蓝点）在帧与帧之间的呈现不一致。而在 DIAMOND 中，这些不一致现象都没有出现。在 *Breakout* 中，当一块红砖被击碎时，分数甚至会可靠地更新 +7[3]。

从图 5 中我们可以看到，与 IRIS 所生成的轨迹相比，DIAMOND 所生成的轨迹视觉质量更高，也更忠实于真实环境。特别是，IRIS 生成的轨迹包含帧与帧之间的视觉不一致（用白色方框标出），例如敌人显示为奖励，反之亦然。这些不一致在生成的图像中可能只代表几个像素，但却会对强化学习产生重大影响。例如，由于代理通常应以奖励为目标并避开敌人，这些微小的视觉差异会增加学习最优策略的难度。

如表 1 所示，视觉细节一致性的改善通常体现在代理在这些游戏中表现更佳。由于这些方法的代理部分是相似的，因此这种改进很可能归因于世界模型。

最后，我们注意到这一改进并不仅仅是计算量增加的结果。两个世界模型都以相同的分辨率（64× 64）渲染帧，DIAMOND 每帧只需要 3 个 NFE，而 IRIS 每帧需要 16 个 NFE。如附录 H 所述，DIAMOND 的参数明显少于 IRIS，训练所需的时间也更短。

6 将扩散世界模型扩展到《*反恐精英：全球攻势*》(4)*全球攻势*[4]

为了研究 DIAMOND 的扩散世界模型学习对更复杂的 3D 环境建模的能力，我们在流行的视频游戏《*反恐精英：全球攻势（CS:GO）*》的静态数据上对世界模型进行了单独训练：*全球攻势*》（CS:GO）的静态数据上单独训练世界模型。我们使用的*在线*数据集是 Pearce 和 Zhu（2022 年）在地图 *Dust II* 中以 16Hz 捕捉到的 550 万帧（95 小时）人类在线游戏数据。我们随机抽取 0.5 百万帧（相当于 500 集或 8 小时）进行测试，并使用剩余的 5 百万帧（87 小时）进行训练。这些实验不涉及强化学习代理或在线数据收集。

[3] https://en.wikipedia.org/wiki/Breakout_(video_game)#Gameplay

[4]本节是在 NeurIPS 验收之后，根据社区对后来 CS:GO 实验的兴趣而添加的。

为了降低计算成本，我们将世界建模的分辨率从 (280× 150) 降低到 (56× 30)。然后，我们引入第二个较小的扩散模型作为上采样器，以改进在原始分辨率下生成的图像（Saharia 等人，2022b）。我们扩展了 U-Net 的通道，使参数数量从 Atari 模型的 400 万增加到 CS:GO 模型的 3.81 亿（包括上采样器的 5100 万）。综合模型在 RTX 4090 上进行了 12 天的训练。

最后，我们引入了随机取样，并将上采样器的去噪步数增加到 10 步，我们发现这提高了生成的视觉质量，同时保持了动态模型不变（特别是仍然只使用 3 个去噪步数）。这样就能在视觉质量和推理成本之间做出合理权衡，模型在 RTX 3090 上以 10Hz 的频率运行。该模型的典型世代见下图 6。

图 6：人们在 DIAMOND 的扩散世界模型中使用键盘和鼠标进行游戏时捕捉到的图像。该模型是在 87 小时的静态《*反恐精英：全球攻势（CS：GO）》游戏（Pearce 和 Zhu，2022 年）*中训练出来的：*全球攻势*》（CS:GO）游戏（Pearce 和 Zhu，2022 年）中进行了训练，从而为流行的游戏地图 "*尘埃 II* "制作了一个交互式神经游戏引擎。最佳观看视频：https://diamond-wm.github.io。

我们发现，该模型能够在数百个时间步长内生成稳定的轨迹，但在地图上不常出现的区域更容易偏离分布。由于模型的内存有限，接近墙壁或失去可见度可能会导致模型忘记当前状态，转而生成新的武器或地图区域。有趣的是，我们发现模型通过概括跳跃对场景几何形状的影响，错误地启用了连续跳跃，因为多次跳跃在训练游戏中出现的频率并不足以让模型学会应该忽略半空中的跳跃。我们希望通过扩展模型和数据来解决上述许多限制，但模型的记忆力除外。对《CS：GO》世界模型能力的定量测量以及解决这些局限性的尝试将留待今后的工作中进行。

7 相关工作

世界模型。Ha 和 Schmidhuber（2018 年）提出了神经网络世界模型想象中的强化学习（RL）理念。SimPLe（Kaiser等人，2019）将世界模型应用于Atari，并引入了Atari 100k基准，重点关注采样效率。Dreamer（Hafner等人，2020年）从递归状态空间模型（RSSM）的潜空间引入了RL。DreamerV2（Hafner 等人，2021 年）证明了使用离散潜变量有助于减少复合误差，而 DreamerV3（Hafner 等人，2023 年）则能够在具有固定超参数的广泛领域中实现人类水平的性能。TWM（Robine等人，2023年）将DreamerV2的RSSM改编为使用变压器架构，而STORM（Zhang等人，2023年）以类似的方式改编了DreamerV3，但采用了不同的标记化方法。另外，IRIS（Micheli 等人，2023 年）使用离散自动编码器构建图像标记语言，并使用自回归变换器随时间推移合成这些标记。

生成视觉模型。这些世界模型与图像生成模型之间存在相似之处，这表明生成视觉模型的发展可为世界建模带来益处。随着变换器在自然语言处理领域的兴起（Vaswani et al、

2017；Devlin 等人，2018；Radford 等人，2019）、VQGAN（Esser 等人，2021）和 DALL-E（Ramesh 等人，2021）通过离散自动编码器将图像转换为离散令牌（Van Den Oord 等人，2017），并利用自回归变换器的序列建模能力建立强大的文本到图像生成模型。与此同时，扩散模型（Sohl-Dickstein 等人，2015 年；Ho 等人，2020 年；Song 等人，2020 年）获得了广泛的关注（Dhariwal 和 Nichol，2021 年；Rombach 等人，2022 年），并已成为高分辨率图像生成的主流范式（Saharia 等人，2022a 年；Ramesh 等人，2022 年；Podell 等人，2023 年）。

视频生成方法的最新发展也呈现出同样的趋势。VideoGPT （Yan 等人，2021 年）通过将离散自动编码器与自回归变换器相结合，提供了一种最小视频生成架构。Godiva（Wu 等人，2021 年）实现了文本调节，具有良好的通用性。Phenaki （Villegas 等人，2023 年）可通过顺序提示调节生成任意长度的视频。TECO（Yan 等人，2023 年）通过使用 MaskGit（Chang 等人，2022 年）改进了自回归建模，并通过压缩输入序列嵌入实现了更长的时间依赖性。扩散模型在使用三维 U-Nets 生成视频方面也出现了回潮，以提供高质量但持续时间较短的视频（Singer 等人，2023 年；Bar-Tal 等人，2024 年）。最近，基于变换器的扩散模型，如 DiT（Peebles 和 Xie，2023 年）和 Sora（Brooks 等人，2024 年），已分别在图像和视频生成方面显示出更好的可扩展性。

扩散强化学习。人们对将扩散模型与强化学习相结合也很感兴趣。这包括利用扩散模型的灵活性作为策略（Wang 等人，2022 年；Ajay 等人，2022 年；Pearce 等人，2023 年）、作为规划器（Janner 等 人，2022 年；Liang 等人，2023 年）、作为奖励模型（Nuti 等人，2023 年）以及在离线 RL 中用于数据增强的轨迹建模（Lu 等人，2023 年；Ding 等人，2024 年；Jackson 等人，2024 年）。DIAMOND 首次将扩散模型用作在线想象学习的世界模型。

生成式游戏引擎。最近，完全基于神经网络运行的可玩游戏范围不断扩大。*GameGAN*（Kim等人，2020年）使用GAN（Goodfellow等人，2014年）学习游戏生成模型，而Bamford和Lucas（2020年）则使用神经GPU（Kaiser和Sutskever，2015年）。同时进行的工作包括 *Genie*（布鲁斯等人，2024 年）和 *GameNGen*（瓦列夫斯基等人，2024 年），前者可根据图像提示生成可玩的平台游戏环境，后者同样利用扩散模型获得高分辨率的游戏 DOOM 模拟器，但规模更大。

8 局限性

我们为未来研究指出了我们工作的三大局限性。首先，我们的主要评估侧重于离散控制环境，而将 DIAMOND 应用于连续领域可能会带来更多启发。其次，使用帧堆叠进行调节是提供过去观测记忆的最基本机制。使用 Peebles 和 Xie（2023 年）等方法将环境时间的自回归变压器整合起来，可以实现更长期的记忆和更好的可扩展性。我们在附录 M 中对潜在的交叉注意力架构进行了初步研究，但在早期实验中发现帧堆叠更为有效。第三，我们将奖励/终结预测与扩散模型的潜在整合留待未来工作进行，因为结合这些目标并从扩散模型中提取表征并非易事（Luo 等人，2023 年；Xu 等人，2023 年），这将使我们的世界模型变得不必要的复杂。

9 结论和更广泛的影响

我们介绍了 DIAMOND，这是一种在扩散世界模型中训练的强化学习代理。我们解释了我们在设计上所做的关键选择，以适应世界模型的扩散，并使我们的世界模型在较长的时间跨度内保持稳定，同时减少去噪步骤的数量。在久负盛名的 Atari 100k 基准测试中，DIAMOND 获得了 1.46 的人类标准化平均分；这是完全在世界模型中训练的代理中的新最佳成绩。我们分析了我们在某些游戏中提高的性能，发现这可能是由于对关键视觉细节进行了更好的建模。通过在静态*《反恐精英：全球攻势》*游戏中进行训练，我们进一步证明了 DIAMOND 的扩散世界模型可以成功地模拟 3D 环境并用作实时神经游戏引擎*：全球攻势》*游戏的训练，进一步证明了 DIAMOND *的扩散世界模型可以成功地模拟 3D* 环境并作为实时神经游戏引擎。

世界模型是解决与真实世界中训练代理相关的样本效率和安全性问题的一个很有前途的方向。然而，世界模型的不完善可能会导致次优或意想不到的代理行为。我们希望，开发更加忠实和互动的世界模型将有助于进一步降低这些风险。