arXiv:2602.02454v1 [cs.RO] 2 Feb 2026

World-Gymnast：在世界模型中通过强化学习训练机器人

Ansh Kumar Sharma [1] **Yixiang Sun** [1] **Ninghao Lu** [2] **Yunzhe Zhang** [1] **Jiarao Liu** [3] **Sherry Yang** [1]

world-gymnast.github.io

摘要

机器人通过与物理世界交互进行学习，根本上受限于物理交互的成本。两种替代方案——基于专家示范进行监督微调（SFT），以及在基于软件的模拟器中进行强化学习（RL）——分别受限于可用专家数据的数量，以及操作任务中的仿真到现实差距。随着近期基于真实世界视频-动作数据学习的世界模型兴起，我们提出一个问题：在世界模型中训练策略，能否比监督学习或软件模拟更有效地提升真实机器人性能？我们提出了 World-Gymnast：在动作条件视频世界模型中对视觉-语言-动作（VLA）策略进行 rollout，并使用视觉语言模型（VLM）为 rollout 结果提供奖励，从而进行强化学习微调。在 Bridge 机器人配置中，World-Gymnast 的表现最高可达 SFT 的 18 倍，并最高可达软件模拟器的 2 倍。更重要的是，World-Gymnast 展示了借助世界模型进行强化学习的诸多潜力，包括在世界模型中针对多样化的语言指令和新场景进行训练、在新场景中进行测试时训练，以及在线迭代式世界模型与策略改进。我们的结果表明，在云端学习世界模型并训练机器人策略，或许是弥合演示中能够完成任务的机器人与能在任何人的家中工作的机器人之间差距的关键。

1. 引言

在现实世界中通过试错学习的机器人面临着一个固有的限制：物理交互成本高昂（Kor- mushev 等人，2013年）。每次依赖于在硬件上执行动作的策略更新都会耗费操作人员的时间、造成设备磨损，并加剧安全隐患（Brunke 等人，2022），尤其是在操作任务中，学习初期失败十分常见。这种成本构成了通过交互扩展机器人学习的根本瓶颈。因此，许多真实机器人系统依赖于能够减少或取代机器人端探索的替代方案（Matas 等人，2018年；Schaal，1999）。

一种替代方案是基于专家示范进行监督学习（SFT），让机器人模仿通过遥操作或脚本控制器收集的轨迹（Ross 等人，2011）。然而，示范数据往往只涵盖长尾情境中很窄的一部分（Hu 等人，2024年），也很少让机器人接触到稳健部署所需的复合错误和恢复行为（Lu 等人，2022）。第二种替代方案是在软件模拟器中进行强化学习（RL）（Zhao 等人，2020）。然而，针对每一种新场景构建软件模拟器的成本都很高。此外，模拟器通常存在仿真到现实差距，例如视觉特征与真实世界图像不同（Salvato 等人，2021）。

近期研究表明，利用真实机器人数据学习得到的世界模型能够近似真实机器人执行的结果（Yang 等人，2023年；Quevedo 等人，2025年；Guo 等人，2025年；Tseng 等人，2025年；Li 等人，2025b）。这些模型旨在预测视觉世界如何随机器人的动作而演变，实际上充当了一个基于真实世界数据学习得到的动作条件视频模拟器。与软件模拟器相比，基于视频的世界模型有望弥合视觉差距，并泛化到新颖的初始帧。然而，由于存在幻觉，目前尚不清楚视频世界模型是否比传统模拟器具有更真实的物理特性（Yang 等人，2024年）。尽管物理真实性难以评估，我们转而探究端到端问题：在学习得到的世界模型中训练机器人策略，能否提升真实机器人性能？

[1]纽约大学 [2]纽约大学上海分校 [3]加州大学伯克利分校。通讯作者：Ansh Kumar Sharma<as20482@nyu.edu>、SherryYang <sherryyang@nyu.edu>。

预印本。*2026*年*2*月*3*日。

相比监督微调或在传统模拟器中进行强化学习?

本文从大型视觉语言动作（VLA）策略的角度探讨这一问题。这类策略将图像和语言指令映射为机器人动作（Brohan 等人，2022；Kim 等人，2024年；Black 等人）。具体而言，我们提出了 World-Gymnast，这是一种利用世界模型对视觉语言动作策略进行强化学习微调的训练框架（图1）。具体来说，World-Gymnast 使用类似于Quevedo 等人2025年的动作条件视频生成模型作为世界模型，使策略能够基于从视觉语言动作策略采样的动作序列生成想象轨迹，并使用视觉语言模型（VLM）从预测视频帧中计算奖励。随后利用这些奖励对视觉语言动作策略进行策略梯度更新。更重要的是，World-Gymnast 为基于世界模型进行强化学习训练开辟了许多引人入胜的可能性，包括：（i）从任意图像帧开始进行强化学习训练；（ii）在新颖的初始帧上进行测试时训练；以及（iii）在线迭代式世界模型与策略改进。

我们通过面向公众开放的自动化真实机器人评估平台 AutoEval（Zhou 等人，2025年），在 Bridge 机器人平台上评估了 World-Gymnast。在 AutoEval 的一系列操作任务中，我们证明，World-Gymnast 的性能显著优于使用原始 Bridge 数据进行监督微调（Walke 等人，2023年），也优于在 SIMPLER（Li 等人，2024年）中进行的强化学习；SIMPLER 是通过真实到仿真技术为 Bridge 创建的软件模拟器。此外，由于世界模型只需一张初始帧即可执行 rollout，我们展示了世界模型一些颇具吸引力的用法，包括基于新颖的语言指令和加入干扰物体的初始帧进行训练、利用真实机器人帧进行测试时训练，以及迭代改进世界模型和策略；所有这些方法都有助于提升真实机器人性能。

2. 预备知识

本节将定义符号并回顾基于模型的强化学习。随后，我们将讨论在基于模型的强化学习框架下，基础世界模型和视觉语言模型如何充当通用动力学模型和奖励模型。

马尔可夫决策过程。我们考虑一个多任务、有限时域、部分可观测的马尔可夫决策过程（POMDP）（Puterman，2014；Kaelbling 等人，1995），由 $\mathcal{M} = (S, A, O, G, R, T, \mathcal{E}, H)$ 指定，包含状态、动作、观测和任务空间，以及奖励、状态转移函数、观测发射函数和时域长度。策略π 与环境交互，执行一个任务，其起始状态为 $g, s_0 \sim G, o_0 \sim \mathcal{E}(s_0)$，产生关于 A 的分布$\pi(\cdot|o_t, g)$，从中采样动作 a_t 并在每一步 $t \in [0, H]$ 应用于环境。环境产生标量奖励 $r_t = R(s_t, g)$，并转移至新状态 $s_{t+1} \sim T(s_t, a_t)$，并发射新的观测 $o_{t+1} \sim \mathcal{E}(s_{t+1})$。

策略 π 的价值可以定义为未来奖励的期望总和：

$$\rho(\pi) = \mathbb{E}[R(s_H, g) | s_0, g \sim G, o_t \sim \mathcal{E}(s_t), a_t \sim \pi(o_t, g), s_{t+1} \sim T(s_t, a_t) \ \ \forall t \in [0, H]]. \tag{1}$$

基于基础模型的强化学习。强化学习（Sutton 等，1998）旨在通过策略与环境之间反复试验和交互，使 $\rho(\pi)$ 最大化。基于模型的强化学习（Doya 等，2002）考虑这样一种情形： T 和 R 未知，需要根据从环境中采集的样本进行估计；这些样本可以是记录先前交互 $D = \{\tau_i = g, s_0, o_0, a_0, ..., s_H, o_H, r_H\}$的离线数据集。考虑到真实世界系统的特点，例如基于图像的观测ˆ以及较高的控制频率，所学习的模型 $T(\cdot|o, a)$通常可以接收一系列先前的图像观测ˆ ˆ和一系列后续动作。在根据数据估计出 T和 R 后，策略便可在学习到的模型中执行 rollout

$$\hat{\rho}(\pi) = \mathbb{E}[\hat{R}([o_0, ..., o_H], g) | s_0, g \sim G, \mathbf{a} \sim \pi(\mathbf{o}, g), \mathbf{o}' \sim \hat{T}(\mathbf{o}, \mathbf{a}), \mathbf{o} = \mathbf{o}']. \tag{2}$$

ˆ 近期研究表明， T可以使用动作条件视频生成模型（世界模型）进行参数化，而 R 可以使用视觉语言模型（VLM）进行参数化。

策略梯度方法（Williams，1992）估计方程2 关于策略 π的梯度，并通过梯度上升直接最大化 $\rho(\pi)$ 。最常用的梯度估计量形式如下：

$$\nabla_\theta \rho(\pi_\theta) = E_{\tau \sim \pi, T} \left[\sum_{t=0}^{H} \gamma^t \nabla_\theta \log \pi_\theta(o_t, g) \hat{A}(o_t, a_t) \right], \tag{3}$$

ˆ其中 A 是一个优势函数，可以通过从 π中获取蒙特卡洛回报来单独估计， T， R (Williams，1992）。采用基于模型的策略梯度时，这些优势函数ˆ ˆ可以根据从 π中获取的蒙特卡洛样本进行估计， T， R。

3. 基于世界模型的强化学习

本节将介绍 World-Gymnast 在第 3.1 节中使用的强化学习算法。随后，我们将在第 3.2 节介绍新兴的训练场景，例如使用分布外（OOD）语言和图像进行训练，以及在第 3.3 节介绍测试时训练。最后，我们将说明如何将 World-Gymnast 与 Dyna 等经典算法（ Sutton， 1991 ）相结合，以实现世界模型与策略的在线迭代改进。

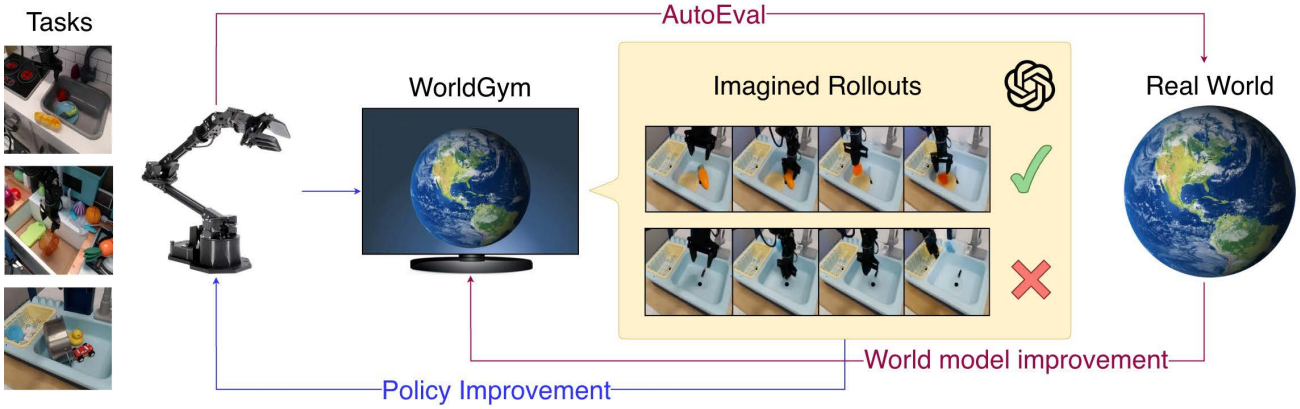


图*1.* **World-Gymnast 概述。** 策略在由初始帧和语言指令指定的任务上进行训练。训练期间，策略输出动作，这些动作随后被传入世界模型（WorldGym（Quevedo 等人，2025年）），由其生成想象轨迹。然后，这些轨迹被传入视觉语言模型，后者返回二元任务完成奖励。该奖励用于更新策略。训练完成后，我们使用 AutoEval（Zhou 等人，2025年）设置，在真实机器人上评估策略。AutoEval 得到的真实世界轨迹（帧-动作序列）还可用于进一步改进特定环境下的世界模型。

3.1. 基于世界模型轨迹的基于模型的GRPO

为了优化策略 π_θ （公式（3）），World-ˆGymnast 使用 T Quevedo 等人 （2025年）提出的已学习世界模型。我们采用组相对策略优化（GRPO）（Shao 等人，2024年），这是一种策略梯度算法，ˆ通过基于组的得分归一化 A 进行估计。

对于给定的任务指令 g 和初始观测o_0，我们通过在世界模型 ˆT中运行策略 π_θ，生成一组 K 条相互独立的轨迹$\{\tau_1, \ldots, \tau_K\}$。具体而言，对于第 k条轨迹，策略采样一个动作 $a_{t,k} \sim \pi_\theta(\cdot|o_{t,k}, g)$，世界模型则预测下一观测 $o_{t+1,k} \sim$ ˆ$T(o_{t,k}, a_{t,k})$。这一过程持续进行，直至达到时域 H，从而得到一条轨迹$\tau_k = (o_{0,k}, a_{0,k}, \ldots, o_{H,k})$。rollout 完成后，ˆ我们使用视觉语言模型 R 为每条轨迹 $r_k = R(\tau_k, g)$分配二元任务完成ˆ奖励。为计算优势函数，我们将这组 K 个输出视为基线，并计算组内奖励的均值和标准差：

$$\mu = \frac{1}{K}\sum_{k=1}^{K} r_k, \quad \sigma = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(r_k - \mu)^2}. \tag{4}$$

随后通过归一化计算第 k 条轨迹的优势函数：

$$\hat{A}_k = \frac{r_k - \mu}{\sigma + \epsilon}, \tag{5}$$

其中，ϵ 是一个用于数值稳定性的较小常数。我们将轨迹级优势函数分配给该轨迹中的每个时间步 t。也就是说，对于所有 $t \in [0, H-1]$，ˆ$A_{t,k}$ = ˆA_k。最后，我们使用基于所计算优势函数进行裁剪的PPO式目标函数来优化策略 π_θ。

损失函数定义如下：

$$\mathcal{J}(\theta) = \mathbb{E}_{g,o_0\sim\mathcal{D}}\Bigg[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H}\sum_{t=0}^{H-1} \min\Big(r_{t,k}(\theta)\hat{A}_k, \text{clip}(r_{t,k}(\theta), 1-\epsilon_{low}, 1+\epsilon_{high})\hat{A}_k\Big)\Bigg], \tag{6}$$

其中，$r_{t,k}(\theta) = \frac{\pi_\theta(a_{t,k}|o_{t,k},g)}{\pi_{\theta_{old}}(a_{t,k}|o_{t,k},g)}$ 表示概率比。

遵循Li 等人(2025a) 使用强化学习进行视觉-语言-动作训练的成功方案，我们采用了其中的一些技巧：1）舍弃 KL 惩罚项；2）通过动态采样过滤掉奖励没有差异的组；3）在 GRPO 中采用更高的裁剪上限；4）在 rollout 期间使用更高的温度采样动作。这些技巧有助于稳定训练，并改善训练过程中的探索。

3.2. 世界模型中的多样化训练场景

在多样化数据集上预训练的世界模型，使我们能够仅凭图像和语言指令生成多样化的训练配置（例如任务和初始观测）。与为每种新配置搭建基于软件的仿真相比，这种方式更加灵活。现在，我们将探索 World-Gymnast 支持的多样化配置下策略训练的多种可能性。

从任意帧开始训练。我们可以使用强化学习训练策略，将任何足够接近世界模型训练分布的帧用作初始观测 o_0,ˆ，随后运行策略 π、世界模型 T和ˆ奖励模型 R，以便针对这一初始配置提供学习信号。这种灵活性大幅增加了

可用于强化学习的有效训练数据量；相比之下，监督微调受限于专家示范的数量。从任意帧开始进行强化学习训练，还能让策略学会恢复行为，从而提高策略的鲁棒性。

针对新颖语言指令进行训练。 通过修改与同一初始帧关联的语言指令，可以进一步扩充训练数据。例如，我们可以向视觉语言模型提供一帧初始画面，并让它提出机器人从该初始帧出发可以执行的合理任务。接着，我们将这些合理任务作为语言指令输入视觉语言动作策略，评估该策略在分布外语言任务上的性能，并通过强化学习进一步提升策略在这些任务上的表现。这样，策略便能针对新任务进行训练，并与环境中先前已存在但未曾明确交互过的物体进行交互。先前关于策略评估的研究表明，经过预训练的视觉语言动作策略往往无法遵循分布外语言指令（Quevedo 等人，2025年）。通过在世界模型中进行强化学习后训练，我们可以克服视觉语言动作策略的这些局限。

利用干扰物进行训练。 为了提升策略对无关视觉杂乱内容的鲁棒性，我们利用 Nano Banana（Google，2025年）等图像编辑工具，在输入图像帧中合成额外物体作为干扰物。使用多样化的干扰物体训练策略，有助于提升策略在实际部署时面对这类干扰物体的鲁棒性，并改善其在杂乱场景中的性能。这有助于弥合只能在演示中工作的机器人与能够在任何人家中杂乱环境里工作的机器人之间的差距。

3.3. 从新帧开始进行测试时训练

由于 World-Gymnast 允许策略仅从初始帧开始进行 rollout，因此在测试时向策略提供新帧后，策略可以通过从该测试帧开始在世界模型中进行强化学习训练，在计算资源和策略性能提升之间进行权衡。这样，策略便能迅速适应新场景，同时避免收集真实世界交互数据所带来的成本和风险。

3.4. 世界模型与策略的迭代改进

当策略 rollout 过程中遇到的视觉观测与世界模型的原始训练分布偏离过大时，直接在世界模型中对策略进行 rollout 可能会导致建模误差不断累积。受经典 Dyna 风格算法（Sutton，1991）的启发，World-Gymnast 允许一种迭代训练流程，其中策略与世界模型交替优化。具体而言，当前策略可以进行 rollout（通过推理时扩展或测试时训练，并使用将世界模型用作奖励函数），以收集新的环境交互数据，随后将其纳入以进一步微调世界模型。更新后的世界模型接着用于生成更优的想象轨迹，供策略优化使用。这一数据飞轮使世界模型能够逐步适应策略所诱导的状态分布，同时让策略受益于世界模型日益准确的长时域预测。

4. 实验

现在，我们评估在 World-Gymnast 中训练的策略的性能。我们在第4.1节介绍实验设置，随后在第4.2节将其与使用软件模拟器和监督微调训练的策略进行比较。接着，我们将在第4.3节展示 World-Gymnast 在支持基于带有干扰物的图像进行多样化训练、处理新颖的语言指令以及扩展强化学习任务数量方面的能力。最后，我们将在第4.4节和第4.5节评估测试时扩展，以及策略与世界模型的迭代改进。

4.1. 实验设置

任务与流程。 我们使用 Kim 等人(2024年) 所采用的精选评估数据集来评估 World- Gymnast 的有效性。该数据集遵循 BridgeData V2（Walke 等人，2023年）设置，使用 WidowX 机器人，旨在测试策略在视觉、运动、物理和语义变化下的泛化能力，以及在 17 个任务上的语言落地能力（附录 C）。

我们进一步利用 World-Gymnast 所支持的数据扩展能力，包括图像编辑、语言增强和新颖的任务设置，以提升训练所得强化学习策略的泛化能力。要使用 World-Gymnast 进行训练，一个任务只需要初始帧和语言指令。在训练过程中，World-Gymnast 在 WorldGym 中将策略 rollout 最多 40 步，并由 GPT-4o（Hurst 等人，2024年）为该 rollout 分配二元任务完成奖励。

训练完成后，我们在 WorldGym（Quevedo 等人，2025年）上评估策略性能，以估计真实机器人性能并确保策略在测试中的安全性；除非另有说明，否则遵循其默认配置（附录A.4）。最后，我们在 AutoEval（Zhou 等人，2025年）上运行该策略。AutoEval 是一个真实机器人评估框架，目前支持 2 种设置下的 4 个任务。AutoEval 对每个策略—任务组合进行 10 次试验；我们重复该评估 5 次，以估计标准误。

基础模型。 成功进行强化学习微调需要一个具备相当能力的初始策略。为此，我们使用 OpenVLA-OFT（Kim 等人，2025年）作为基础模型。OpenVLA-OFT

表1。 **AutoEval 真实机器人成功率** (Zhou 等人，2025年)，比较 World-Gymnast 与在软件模拟器 SIMPLER 中运行强化学习（Li 等人，2024年）。在 4 项任务中的 3 项上，基于世界模型的强化学习在真实机器人成功率方面显著优于在模拟器中进行的强化学习。

Task	SIMPLER	World-Gymnast
打开抽屉	$34 \pm 7\%$	$58 \pm 4\%$
关闭抽屉	$74 \pm 5\%$	$62 \pm 6\%$
把茄子放进蓝色水槽	$32 \pm 10\%$	$72 \pm 10\%$
把茄子放入黄色篮子	$40 \pm 10\%$	$78 \pm 2\%$

提供了一套优化的微调方案，可在 OpenVLA（Kim 等人，2024年）基础上进一步构建；OpenVLA 最初是在 Open X-Embodiment 数据集（Vuong 等人，2023年）上训练的。我们使用 BridgeData V2（Walke 等人，2023年）对基础模型进行微调。遵循（Li 等人，2025a）中的思路，我们对 OpenVLA-OFT 的官方实现做了几项修改：1）禁用本体感觉和辅助摄像头输入，使其与策略所用的观测空间一致；2）使用 LLAMA-2（Touvron 等人，2023年）的语言模型头作为动作头，替代默认的 L1 损失动作头，从而获得强化学习所必需的动作概率。对于 WorldGym，我们使用了一个在 Open X-Embodiment 数据集上预训练的 6 亿参数变体。

训练细节。 对于强化学习训练，我们使用 4 块 NVIDIA H200 GPU（每块 140GB）进行全参数微调，耗时 1–2 天。我们使用以下训练参数：学习率 $5 \cdot 10^{-6}$，组大小 8，训练批次大小 20，动作块长度 5，裁剪比率（$\epsilon_{high} = 0.28$，$\epsilon_{low} = 0.2$），温度 1.6。更详细的超参数设置见附录 A.4。

4.2. 使用 World-Gymnast 评估强化学习

将 World-Gymnast 与软件模拟器进行比较。我们将 World-Gymnast 与基于传统模拟器的强化学习进行比较。我们选择 SIMPLER（Li 等人，2024年）这一现实到模拟的策略评估框架作为基线，因为它提供了与我们评估中使用的 Bridge机器人和任务最接近的基于模拟器的近似。我们在 SIMPLER 中所有可用的任务上进行训练（附录 B.2），并进一步为 AutoEval 配置纳入数字孪生。在从基础监督微调策略初始化强化学习时，我们观察到，除关闭抽屉外，SIMPLER 中所有任务的任务完成率都很低，尽管策略通常会朝正确方向移动，却无法完全完成任务。因此，使用二元完成奖励会导致奖励方差坍缩，从而无法进行有效的策略更新。为解决这一问题，我们将一次 rollout 的奖励定义为每一步部分奖励之和。关于奖励设计的更多细节见附录B.2。

表2。 **真实机器人任务成功率** World-Gymnast 与监督学习方法的对比。标准误差根据每组 10 次连续回合计算。

Task	SFT	Iter-SFT	World-Gymnast
打开抽屉	$40 \pm 10\%$	$30 \pm 10\%$	$58 \pm 4\%$
关闭抽屉	$62 \pm 6\%$	$60 \pm 8\%$	$62 \pm 6\%$
把茄子放进蓝色水槽	$4 \pm 4\%$	$10 \pm 3\%$	$72 \pm 10\%$
把茄子放进黄色篮子	$8 \pm 4\%$	$17 \pm 17\%$	$78 \pm 2\%$

在除关闭抽屉之外的所有任务上，World-Gymnast 的表现都优于使用 SIMPLER 进行训练的结果，如表 1所示。对于关闭抽屉 任务，基础策略在强化学习之前已经表现得相当不错。值得注意的是，World-Gymnast 的强化学习训练集不包含这些训练集中的任务，而 SIMPLER基线却在完全相同的环境—任务配置（数字孪生）上进行了训练，但该策略对现实世界的迁移表现仍然很差。

将 World-Gymnast 与监督学习进行比较。我们还将 World-Gymnast 与监督微调方法进行比较。遵循 OpenVLA-OFT（Kim 等人，2025年）的方法，我们在 Bridge V2 数据集（Kim 等人，2024年）中的专家轨迹上对 OpenVLA 70亿策略进行微调（Walke 等人，2023年），训练 20k 步。该策略在表 *2* 中记为 SFT，也是我们进行强化学习训练所使用的基础模型。近期的工作（如 Ctrl-World（Guo 等人，2025年））进一步利用世界模型，通过生成由奖励模型筛选的合成回合，将其作为额外的监督信号来改进策略。与 Ctrl-World 类似，我们在世界模型中运行基础 *SFT* 策略，运行步数与 World-Gymnast 进行强化学习训练时相同，并使用视觉语言模型在 OpenVLA评估任务上筛选出成功轨迹。随后，我们使用 Bridge V2 数据与成功的合成回合的混合数据，以 $1:1$ 采样比例进行新一轮监督微调。所得策略在表 *2* 中记为 Iter-SFT，随后通过 AutoEval 在现实世界中未参与训练的任务上进行评估。

如表2所示，与监督学习和 Iter- SFT 相比，World-Gymnast 的性能最佳，在把茄子放进蓝色水槽和把茄子放进黄色篮子这两项任务上的表现分别比基础策略提升了 18 倍和近 10 倍。值得注意的是，Iter-SFT 在较难的任务上略有提升，但在较简单的任务上性能反而下降。一种可能的解释是，强化学习通过主动探索和同策略更新，能够学到泛化性更强的行为。相比之下，迭代式监督微调可能会过拟合合成经验，并且更容易受到世界模型幻觉和不准确的视觉语言模型成功判定的影响。

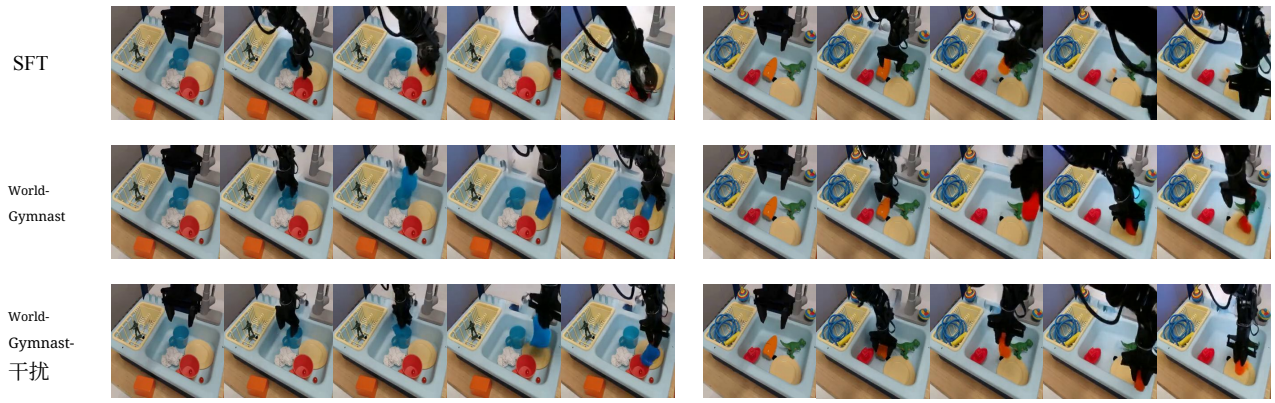


图*2.* **在含干扰物的 WorldGym 中对策略 rollout 进行定性评估**。 我们比较了监督微调、World-Gymnast 和 World-Gymnast-Distract 在视觉干扰下的 rollout 质量。左侧任务是将蓝色杯子放到盘子上，监督微调策略明显拿起了错误的杯子，而两种 World-Gymnast 变体都能正确执行该任务。在右侧任务（将胡萝卜放到盘子上）中，我们可以看到监督微调再次遇到困难，似乎在抓取胡萝卜的同时也抓起了恐龙。两种 World-Gymnast 变体再次成功，但 World-Gymnast-Distract 的抓取和放置动作更好。值得注意的是，即使不完美的世界模型引入了视觉伪影，这些策略仍能有效迁移到真实机器人环境中。

4.3. 评估 World-Gymnast 所提供的多样化设置

评估使用干扰物进行训练。我们使用 Nano Ba- nana（Google，2025年）利用 OpenVLA Bridge 任务套件中已有的帧生成一个新数据集。新数据集加入随机物体，以干扰策略成功完成给定任务。随后，我们将这些新帧纳入训练数据，使用强化学习训练一个新策略（World-Gymnast-Distract）。接着，我们在包含干扰物的留出帧上评估监督微调、World-Gymnast 和 World-Gymnast-Distract 的性能。WorldGym 评估结果表明，World-Gymnast-Distract 的鲁棒性最强，而监督微调最容易受到干扰（图2）。此外，我们还在 WorldGym 中使用原始 OpenVLA 任务评估 World-Gymnast-Distract，并观察到其成功率有所提升（表3）。这表明，加入经过干扰物增强的数据，不仅能提升模型在视觉扰动下的性能，也能改善其在原始任务上的表现。图2 中的定性 rollout 对比进一步表明，在存在视觉干扰时，与监督微调和 World-Gymnast 相比，World-Gymnast-Distract 能更可靠地执行抓取和放置动作，同时即使世界模型产生视觉伪影，也能保持正确的物体定位。

评估使用新颖的语言指令进行训练。
扩展数据规模的另一种方法，是为现有任务补充语言指令。我们通过创建 4 项新任务来检验这种方法，这些任务涉及与场景中已有物体进行新的交互。我们将这些新数据与 OpenVLA 数据集结合，并训练了一个新策略 World-Gymnast-Language。接下来，我们在 OpenVLA 数据的留出划分上评估 World-Gymnast-Language 的性能，发现其性能优于 World-Gymnast（表3）。

表 *3.* **比较不同设置下的强化学习**。 借助世界建模的多样化能力，World-Gymnast 能够在多种设置中显著提高任务成功率。

变体	成功率
SFT	$58 \pm 4\%$
World-Gymnast	$74 \pm 3\%$
World-Gymnast-Distract	$78 \pm 2\%$
World-Gymnast-Language	$81 \pm 1\%$
World-Gymnast-Scaled	$81 \pm 4\%$

这表明，通过在现有帧上引入新颖的语言指令来创建更多任务，可以进一步提升视觉-语言-动作策略的泛化性能。

扩展训练任务数量。 World-Gymnast 的一个优势是能够从任意初始帧开始，在多样化的任务集合上进行训练。为了扩展数据规模，我们从 Bridge 数据集（Walke 等人，2023年）中随机选取了另外 5 项任务。随后，我们将这些任务与 OpenVLA 任务一同用于训练，并将该变体称为 World-Gymnast- Scaled。我们在 WorldGym 中，使用 OpenVLA 留出任务划分评估 World-Gymnast- Scaled 的性能，发现其表现有所提升，相较 World-Gymnast 更佳，如表 3所示。这些结果表明，World-Gymnast 能够有效利用额外的训练任务来提升性能。

4.4. 评估测试时优化

预训练策略往往难以很好地泛化到新的现实世界场景。虽然在线数据收集后再进行微调可以弥补这一差距，但在时间和精力上的成本高得令人难以承受。借助预训练的

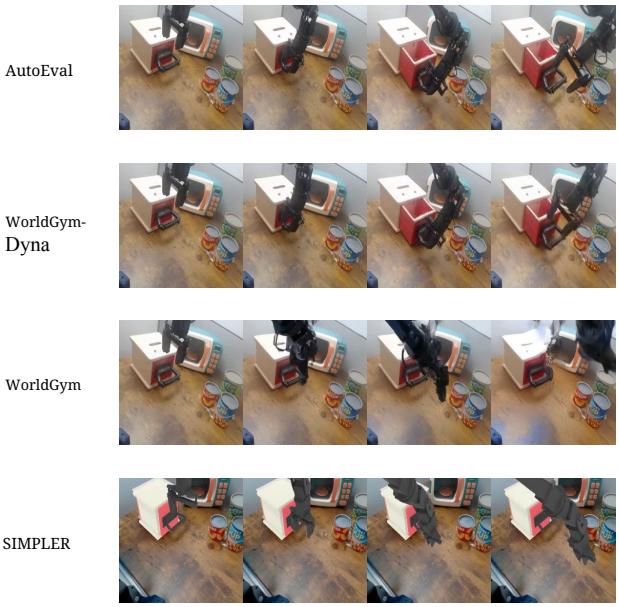


图3. **定性比较：**在真实机器人上执行相同动作序列的 rollout，分别来自 AutoEval（Zhou 等人，2025年）、软件模拟器 SIMPLER（Li 等人，2024年）、WorldGym（Quevedo 等人，2025年）以及采用在线世界模型更新的 World-Gymnast。与 SIMPLER 相比，World-Gymnast 的 rollout 更贴近现实世界，这表明通过 Dyna（Sutton，1991）改进世界模型能够提升 rollout 的质量。

世界模型，我们展示了 World-Gymnast 无需现实世界中的回合，即可通过测试时训练提升基础策略的性能。具体而言，我们仅使用 AutoEval 中 4 个场景的初始观测和任务指令（Zhou 等人，2025年），通过强化学习对基础策略进行微调（详见附录 A.4），并以零样本方式，根据测试帧利用世界模型生成的想象回合进行训练。测试时训练显著提升了现实世界中关闭抽屉任务的性能和鲁棒性，将成功率从$62 \pm 6\%$ 提高到 $100 \pm 0\%$。不过，关闭抽屉任务的测试时训练会导致模型过拟合于这一单一任务，而其他任务的性能普遍下降。在多种任务上进行测试时训练是一个值得探索的未来研究方向。

4.5. 评估世界模型与策略的迭代改进

World-Gymnast 的一项独特优势是，对于初始帧分布在预训练世界模型之外的情况，可以利用现实世界中的回合数据迭代更新世界模型。在使用 AutoEval 进行策略评估时（Zhou 等人，2025年），我们会保存任务指令、观测和动作序列，以改进世界模型。我们对全部 4 项任务分别迭代收集了约 100 条轨迹，并使用这些数据对世界模型进行微调，总计训练 120k 步。图3 给出了一个定性结果，展示了世界模型在执行一系列动作方面的改进，对应的任务是打开抽屉；结果表明，与 SIMPLER（Li 等人，2024年）以及未经在线更新的 WorldGym 相比，其仿真到现实差距小得多。将使用在线数据更新的 Dyna风格世界模型用作强化学习环境后，World-Gymnast 将基础强化学习模型在 AutoEval 中执行关闭抽屉任务的成功率提升至95%。

5. 相关工作

基于模型的强化学习。基于模型的强化学习长期以来一直是强化学习文献中的研究主题：它从先前收集的数据中学习动力学模型，并利用学得的动力学模型进行策略评估和改进（Sutton，1991；Tani，1996；Ljung与Glad，1994；Liu等人，2019；Zhang等人，2021；Yu等人，2020；Hafner等人，2020）。许多基于模型的强化学习研究都聚焦于在低维状态空间中为每个系统学习一个动力学模型，而非在像素空间中学习（Ferns等人，2004；Achille与Soatto，2018；Lesort等人，2018；Castro，2020）。尽管这使建模问题更简单，却限制了不同系统之间的知识共享。随着大型Transformer架构的发展，先学习基于图像的世界模型、再进行强化学习已成为可能（Hafner等人，2020；Chen等人，2022；Seo等人，2022；Micheli等人，2022；Wu等人，2022；Hafner等人，2023），但相关研究大多集中于游戏或模拟环境，这些环境中的视觉数据较为简单且数量充足。我们的工作与现有的基于模型的强化学习研究不同：我们侧重于利用一个基于多种策略和任务的大量数据训练出的单一世界模型，来训练能够处理新颖语言指令和场景的通用视觉语言动作策略。我们还通过与传统模拟器中的强化学习进行比较，直接应对仿真到现实差距，并进一步展示了基于模型的强化学习在模型由真实世界数据学习得到时的价值。

仿真到现实强化学习。基于物理特性的仿真中的强化学习已被广泛采用，以克服现实世界训练的样本低效问题。为弥合现实差距，先前工作在很大程度上依赖于领域随机化（Tobin 等人，2017年；Peng 等人，2018年）——通过改变视觉和物理参数来覆盖现实世界分布——或领域自适应（Bousmalis 等人，2018年；Rao 等人，2020年）。这些策略在运动控制（Tan 等人，2018年；Rudin 等人，2022）和刚体操作（OpenAI 等人，2019年；Handa 等人，2024年）方面取得了显著成功。然而，传统模拟器（Todorov 等人，2012年；Makoviychuk 等人，2021年）在通用操作方面面临可扩展性瓶颈：它们需要显式的物体建模和人工场景工程，并且难以忠实地渲染现实世界多样的视觉纹理和可变形物体动力学。与这些方法不同，World-Gymnast 利用直接从真实世界数据中学习得到的世界模型，有效地绕过了对人工

资产创建和物理参数调优的需求。

用于机器人学习的视频生成。基于视频的机器人学学习（Nair 等人，2022；Bahl 等人，2022；Shao 等人，2021；Chen 等人，2021；Pari 等人，2022）推动了视觉表征学习、目标提取、规划（Finn 和 Levine，2017；Kurutach 等人，2018），以及从专家动作中进行模仿（Fang 等人，2019；Wang 等人，2023；Mani 等人，2024）。近期研究将决策重新表述为文本条件视频生成任务，从而能够基于视频预测进行策略学习（Du 等人，2024；Ko 等人，2023；Wen 等人，2023；Du 等人，2023；Ajay 等人，2024），并使用生成模型模拟智能体与环境交互（Yang 等，2023）。这些研究大多将生成的视频计划用作视觉动作，并训练独立的逆动力学模型，从生成的视频中提取机器人动作。尽管文本生成视频可有效用于长时域规划，但如何自我改进这些将视频生成用作策略的模型，仍不甚明了。我们研究将视频生成仅用作环境，并通过强化学习利用生成的轨迹来提升策略性能这一问题。值得注意的是，原则上，World-Gymnast 可用于改进视觉-语言-动作策略以外的任何策略，包括通过视频生成进行参数化的策略。

使用世界模型进行策略评估。近期研究表明，利用真实机器人数据学习得到的世界模型能够近似真实机器人执行结果，因此可用于评估机器人策略（Quevedo 等人，2025年；Guo et al.，2025年；Tseng etal.，2025年；Li 等人，2025b）。尽管在世界模型中评估策略无疑是世界模型的一项重要应用，但我们聚焦于利用世界模型提升策略性能这一端到端问题，并在真实机器人上检验其效果。

基于视频世界模型的强化学习。我们的工作与Zhu et al.（2025年）最为相似，该研究利用强化学习在视频世界模型中改进策略；但我们侧重于真实世界评估，并使用易于获取的评估设置，包括开源视觉语言动作策略（OpenVLA）、世界模型（WorldGym）和评估平台（AutoEval）。我们还着重探索强化学习在世界模型中的涌现能力，包括从任意初始帧开始、通过新颖的语言指令进行训练、测试时训练，以及世界模型与策略的迭代改进。

6. 结论与局限性

我们介绍了 World-Gymnast，这是一种利用学习得到的世界模型对视觉-语言-动作策略进行微调的强化学习框架。我们表明，现有训练范式（包括监督微调和基于软件模拟器的强化学习）成本高、限制多，且往往会产生泛化能力有限的策略。相比之下，使用世界模型进行策略 rollout，并利用视觉语言模型提供奖励，成本低、可扩展性强，且在分布外场景中更加稳健。

基于世界模型的训练的一项关键优势，是能够从极少的输入生成多样化的训练数据。由于 World-Gymnast 只需要一个初始场景和一段任务描述，它天然支持借助图像编辑、语言变体以及复用新环境中的场景来进行数据增强。这种灵活性进一步使测试时训练、规划以及对策略与世界模型的迭代改进成为可能。

局限性。 World-Gymnast 的一个局限性在于，若初始帧远离世界模型的训练分布，它便无法泛化到任意初始帧。这呼吁未来研究如何在广泛的机器人数据集上预训练机器人世界模型。World-Gymnast 的另一个局限性是依赖预训练视觉语言模型来判定任务成功，这可能产生幻觉，从而导致次优的强化学习训练。探索改进奖励模型的途径，例如训练奖励模型（Lee 等人，2026年），是一个重要的未来方向。此外，利用视觉语言模型的稠密奖励并防止奖励投机，同样是未来工作中颇具前景的方向。

影响声明

本工作介绍了 World-Gymnast，这是一个在生成式世界模型中训练机器人策略的框架。我们的方法有望通过减少对昂贵物理硬件和手动模拟器工程的依赖，降低开发通用机器人的门槛，从而推动机器人研究普及。尽管该方法显著降低了与现实训练相关的物理风险和成本，但我们也承认，依赖生成式视频模型会带来策略利用幻觉物理的风险。因此，尽管我们的结果展示了更出色的仿真到现实迁移能力，但在此类环境中训练的策略在部署到安全关键的现实环境之前，仍应接受严格的安全验证。

7. 致谢

我们谨此感谢 AutoEval（Zhou 等人，2025年）团队，尤其是 Zhiyuan (Paul) Zhou，感谢他们协助开展 AutoEval 实验。我们谨此感谢 Julian Quevedo，感谢他在使用 WorldGym（Quevedo 等人，2025年）代码库方面提供的指导。

参考文献

Achille, A. 和 Soatto, S. 深度学习时代的控制分离原理。《*Annual Review of Control*》，

机器人学与自主系统，1:287–307，2018年。

Ajay, A., Han, S., Du, Y., Li, S., Gupta, A., J aakkola, T., Tenenbaum, J., Kaelbling, L., Srivastava, A., and Agrawal, P. 用于分层规划的组合式基础模型。神经信息处理系统进展，36，2024年。

Bahl, S., Gupta, A., and Pathak, D. 在真实环境中进行人类到机器人的模仿学习。载于机器人学：科学与系统，2022。

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., 等$\pi 0$：一种用于通用机器人控制的视觉-语言-动作流模型。corr, abs/2410.24164，2024年。doi：10.48550。*arXiv* 预印本 *ARXIV.2410.24164*。

Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Kono- lige, K., 等。利用仿真和领域自适应提高深度机器人抓取的效率。载于*2018*年 *IEEE* 国际机器人与自动化会议（*ICRA*），第 4243–4250 页。IEEE，2018年。

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., 等。Rt-1：用于大规模真实世界控制的机器人变换器。*arXiv* 预印本 *arXiv:2212.06817*，2022。

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., 和 Schoellig, A. P. 机器人安全学习：从基于学习的控制到安全强化学习。《控制、机器人与自主系统年度评论》，5(1):411–444，2022。

Castro, P. S. 用于计算确定性马尔可夫决策过程状态相似性的可扩展方法。载于*AAAI*人工智能会议论文集，第34卷，第10069–10076页，2020年。

Chen, A. S., Nair, S. 和 Finn, C. 从“野外”人类视频中学习可泛化的机器人奖励函数。载于机器人学：科学与系统，2021。

Chen, C., Wu, Y.-F., Yoon, J. 和 Ahn, S. Transdreamer：利用Transformer世界模型进行强化学习。*arXiv*预印本 *arXiv:2202.09481*，2022。

Doya, K., Samejima, K., Katagiri, K.-i. 和 Kawato, M. 多模型基于模型的强化学习。神经计算，14(6):1347–1369，2002。

Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter,B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J. B., 等. 视频语言规划。*arXiv* 预印本 *arXiv:2310.10625*，2023年。

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., 和 Abbeel, P. 通过文本引导的视频生成学习通用策略。神经信息处理系统进展，36，2024年。Esser, P., Kulal, S., Blattmann, A., Entezari, R., M¨uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., 等. 用于高分辨率图像合成的整流流Transformer扩展。载于第四十一届国际机器学习大会，2024年。Fang, B., Jia, S., Guo, D., Xu, M., Wen, S., 和 Sun, F. 机器人操作模仿学习综述。国际智能机器人与应用期刊，2019年。

Ferns, N., Panangaden, P. 和 Precup, D. 有限马尔可夫决策过程的度量。载于*UAI*，第4卷，第162–169页，2004。

Finn, C. 和 Levine, S. 用于机器人运动规划的深度视觉前瞻。载于*IEEE*国际机器人与自动化会议，2017年。

Google。Gemini 中的图像编辑迎来重大升级。Google 于 2025年8月26日发表于“The Keyword”的博客文章。网址 https://blog.google/products/gemini/updated-image-editing-model/。多模态生成负责人，Gemini应用；Google DeepMind 的 Gemini Image 产品负责人。

Guo, Y.、Shi, L.X.、Chen, J. 和 Finn, C. Ctrl-world：用于机器人操作的可控生成式世界模型。*arXiv*预印本 *arXiv:2510.10125*，2025年。

Hafner, D.、Lillicrap, T.、Norouzi, M. 和 Ba, J. 使用离散世界模型掌握 Atari。*arXiv*预印本 *arXiv:2010.02193*，2020年。

Hafner, D.、Pasukonis, J.、Ba, J. 与 Lillicrap, T. 通过世界模型掌握多样化领域。*arXiv* 预印本 *arXiv:2301.04104*，2023年。

Handa, A.、Allshire, A.、Makoviychuk, V.、Petrenko, A.、Singh, R.、Liu, J.、Makoviichuk, D.、Wyk, K. V.、Zhurkevich, A.、Sundaralingam, B.、Narang, Y.、Lafleche, J.-F.、Fox, D. 与 State, G. Dextreme：将敏捷手内操作从仿真迁移到现实，2024年。网址 https://arxiv.org/abs/2210.13702。

Hu, Y.、Lin, F.、Sheng, P.、Wen, C.、You, J. 与 Gao, Y. 机器人操作模仿学习中的数据缩放定律。*arXiv* 预印本 *arXiv:2410.18647*，2024年。

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., 等。GPT-4o 系统卡。*arXiv* 预印本 *arXiv:2410.21276*，2024年。

Kaelbling, L. P., Littman, M. L. 和 Cassandra, A. R. 面向人工智能的部分可观测马尔可夫决策过程。载于机器人学不确定性推理国际研讨会，第 146–163 页。Springer，1995年。

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., 等。OpenVLA：一种开源视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2406.09246*，2024年。

Kim, M. J., Finn, C., 和 Liang, P. 视觉-语言-动作模型的微调：优化速度和成功率。*arXiv* 预印本 *arXiv:2502.19645*，2025年。

Ko, P.-C., Mao, J., Du, Y., Sun, S.-H., 和 Tenenbaum, J. B. 通过密集对应关系，从无动作视频中学习动作。*arXiv* 预印本*arXiv:2310.08576*，2023年。

Kormushev, P., Calinon, S., 和 Caldwell, D. G. 机器人学中的强化学习：应用与现实世界挑战。机器人学，2(3):122–148，2013年。

Kurutach, T., Tamar, A., Yang, G., Russell, S. J., 和 Abbeel, P. 使用因果 Info-GAN 学习可规划表示。发表于 神经信息处理系统，2018年。

Lee, T., Wagenmaker, A., Pertsch, K., Liang, P., Levine, S., 和 Finn, C. RoboReward：面向机器人学的通用视觉-语言奖励模型，2026年。URL https://arxiv.org/abs/2601.00675。

Lesort, T., D´ıaz-Rodr´ıguez, N., Goudou, J.-F., 和 Filliat, D. 面向控制的状态表示学习：综述。神经网络，108:379–392，2018年。

Li, H., Zuo, Y., Yu, J., Zhang, Y., Yang, Z., Zhang, K., Zhu, X., Zhang, Y., Chen, T., Cui, G. 等。SimpleVLA-RL：通过强化学习扩展视觉-语言-动作训练。*arXivpreprintarXiv:2509.09674*，2025年a。

Li, X., Hsu, K., Gu, J., Pertsch, K., Mees, O., Walke, H. R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S. 等。在仿真中评估现实世界机器人操控策略。*arXivpreprint arXiv:2405.05941*，2024年。

Li, Y., Zhu, Y., Wen, J., Shen, C. 与 Xu, Y.。Worldeval：将世界模型作为现实世界机器人策略评估器。*arXivpreprintarXiv:2505.19017*，2025年b。

Liu, J., Gu, X. 和 Liu, S. 基于世界模型的强化学习。*arXiv* 预印本 *arXiv:1908.11494*，2019。

Ljung, L. 和 Glad, T. 动态系统建模。Prentice-Hall,Inc.，1994年。

Lu, C., Ball, P. J., Rudner, T. G., Parker-Holder, J., Osborne, M. A., 和 Teh, Y. W. 基于视觉观测的离线强化学习：挑战与机遇。*arXiv* 预印本 *arXiv:2206.04779*，2022年。Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., 和 State, G. Isaac Gym：面向机器人学习的高性能 GPU 物理仿真，2021年。网址：https://arxiv.org/abs/2108.10470。Mani, S., Venkataraman, S., Chandra, A., Rizvi, A., Sirvi, Y., Bhattacharya, S., 和 Hazra, A. Diffclone：利用扩散驱动策略学习增强机器人学中的行为克隆。*arXiv:2401.09243*，2024年。Matas, J., James, S., 和 Davison, A. J. 面向可变形物体操作的仿真到现实强化学习。载于机器人学习会议，第 734–743 页。PMLR，2018年。

Micheli, V., Alonso, E. 和 Fleuret, F. Transformer 是样本效率高的世界模型。*arXiv* 预印本 *arXiv:2209.00588*，2022。

Nair, S., Rajeswaran, A., Kumar, V., Finn, C. 和 Gupta, A. R3M：用于机器人操作的通用视觉表征。发表于机器人学习会议，2022。

OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W. 和 Zhang, L. 使用机器人手解决魔方，2019。网址 https://arxiv.org/abs/1910.07113。

Pari, J., Shafiullah, N. M., Arunachalam, S. P. 和 Pinto, L. 表征学习在视觉模仿中的惊人效果。载于机器人学：科学与系统，2022。

Peng, X. B., Andrychowicz, M., Zaremba, W. 和 Abbeel, P. 通过动力学随机化实现机器人控制的仿真到现实迁移。载于*2018*年 *IEEE* 国际机器人与自动化会议（*ICRA*），第3803–3810页。IEEE，2018年。

Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., Xiao, K., Agrawal, S. 和 Dean, J. 高效扩展 Transformer 推理。机器学习与系统会议论文集，5：606–624，2023年。

Puterman, M.L. 马尔可夫决策过程：离散随机动态规划。约翰·威立父子出版公司，2014。

Quevedo, J., Sharma, A. K., Sun, Y., Suryavanshi, V., Liang, P., and Yang, S. Worldgym：作为策略评估环境的世界模型，2025年。网址 https://arxiv.org/abs/2506.00613。

Rao, K., Harris, C., Irpan, A., Levine, S., Ibarz, J. 和 Khansari, M. RL-CycleGAN：感知强化学习的仿真到现实。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第11157–11166页，2020年。

Ross, S., Gordon, G. 和 Bagnell, D. 将模仿学习和结构化预测归约为无悔在线学习。在第十四届人工智能与统计国际会议论文集，第627–635页。JMLR Workshop and Conference Proceedings，2011年。Rudin, N., Hoeller, D., Reist, P. 和 Hutter, M. 使用大规模并行深度强化学习，在几分钟内学会行走，2022。网址：https://arxiv.org/abs/2109.11978。

Salvato, E., Fenu, G., Medvet, E. 和 Pellegrino, F. A. 跨越现实差距：强化学习中机器人控制器仿真到现实迁移能力的综述。*IEEEAccess*，9:153171–153187，2021年。Schaal, S. 模仿学习是通往人形机器人的道路吗？认知科学趋势，3(6):233–242，1999年。

Seo, Y., Lee, K., James, S. L. 和 Abbeel, P. 利用视频进行无动作预训练的强化学习。在国际机器学习会议，第19561–19579页。PMLR，2022年。Shao, L., Migimatsu, T., Zhang, Q., Yang, K. 和 Bohg, J. Concept2Robot：从指令和人类演示中学习操作概念。*IJRR*，2021年。Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y. 等。DeepSeekMath：推动开放语言模型数学推理能力的极限。*arXiv*预印本 *arXiv:2402.03300*，2024年。Sutton, R. S. Dyna：一种集学习、规划和反应于一体的架构。*ACMSigart Bulletin*，2(4):160–163，1991年。Sutton, R. S., Barto, A. G. 等。强化学习：导论，第1卷。麻省理工学院出版社，剑桥，1998年。Tan, J., Zhang, T., Coumans, E., Iscen, A., Bai, Y., Hafner, D., Bohez, S. 和 Vanhoucke, V. 仿真到现实：学习四足机器人的敏捷运动控制，2018年。网址 https://arxiv.org/abs/1804.10332。

Tani, J. 从动力系统视角出发的移动机器人导航基于模型的学习。*IEEE*《系统、人和控制论汇刊》*B*辑（控制论），26(3):421–436，1996。Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., 和 Abbeel, P. 通过领域随机化将深度神经网络从仿真迁移到现实世界。载于*2017*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第23–30页。IEEE，2017。Todorov, E., Erez, T., 和 Tassa, Y. Mujoco：用于基于模型的控制的物理引擎。载于*IROS* Todorov 等人（2012），第5026–5033页。ISBN 978-1-4673- 1737-5。网址：https://dblp.uni-trier.de/db/conf/iros/iros2012.html#TodorovET12。Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. 等。Llama 2：开放式基础模型与微调聊天模型。*arXiv*预印本 *arXiv:2307.09288*，2023年。Tseng, W.-C., Gu, J., Zhang, Q., Mao, H., Liu, M.- Y., Shkurti, F., 和 Yen-Chen, L. 利用视频世界模型实现可扩展的策略评估。*arXiv*预印本 *arXiv:2511.11520*，2025年。Vuong, Q., Levine, S., Walke, H. R., Pertsch, K., Singh, A., Doshi, R., Xu, C., Luo, J., Tan, L., Shah, D. 等。Open X-Embodiment：机器人学习数据集与RT-X模型。载于迈向通用机器人：可扩展技能习得的学习范式@*CoRL2023*，2023年。Walke, H. R., Black, K., Zhao, T. Z., Vuong, Q., Zheng, C., Hansen-Estruch, P., He, A. W., Myers, V., Kim, M. J., Du, M. 等。BridgeData V2：面向大规模机器人学习的数据集。载于机器人学习会议，第1723–1736页。PMLR，2023年。Wang, H.-C., Chen, S.-F., 和 Sun, S.-H. 扩散模型增强的行为克隆。*arXiv:2302.13335*，2023年。Wen, C., Lin, X., So, J., Chen, K., Dou, Q., Gao, Y., 和 Abbeel, P. 用于策略学习的任意点轨迹建模。*arXiv*预印本 *arXiv:2401.00025*，2023年。Williams, R. J. 用于联结主义强化学习的简单统计梯度跟随算法。机器学习，8(3):229–256，1992年。Wu, Z., Dvornik, N., Greff, K., Kipf, T., 和 Garg, A. Slotformer：利用以物体为中心的模型进行无监督视觉动力学仿真。*arXiv*预印本 *arXiv:2210.05861*，2022年。

Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuur- mans, D., 和 Abbeel, P. 学习交互式现实世界模拟器。*arXiv*预印本 *arXiv:2310.06114*，2023年。Yang, S., Walker, J., Parker-Holder, J., Du, Y., Bruce, J., Barreto, A., Abbeel, P., 和 Schuur mans, D. 视频作为现实世界决策的新语言。*arXiv*预印本 *arXiv:2402.17139*，2024年。Yu, T., Tho mas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., 和 Ma, T. MOPO：基于模型的离线策略优化。载于神经信息处理系统进展，2020年。Zhang, M. R., Paine, T., Nachum, O., Paduraru, C., Tucker, G., ziyu wang, 和 Norouzi, M. 用于离线策略评估与优化的自回归动力学模型。载于国际学习表征会议，2021年。Zhao, W., Queralta, J. P., 和 Westerlund, T. 机器人领域深度强化学习的仿真到现实迁移：综述。载于*2020*年*IEEE*计算智能专题研讨会系列（*SSCI*），第737–744页。IEEE，2020年。Zhou, Z., Atreya, P., Tan, Y. L., Pertsch, K., 和 Levine, S. AutoEval：在现实世界中自主评估通用机器人操控策略。*arXiv*预印本 *arXiv:2503.24278*，2025年。Zhu, F., Yan, Z., Hong, Z., Shou, Q., Ma, X., 和 Guo, S. Wmpo：用于视觉-语言-动作模型的世界模型策略优化。*arXiv*预印本 *arXiv:2511.09515*，2025年。

算法 1 WorldGym 使用时间 KV 缓存进行推理。

初始化：
$x \leftarrow x_0$，当前帧 $\leftarrow 0$，块 $= 1$，最大帧数 ▷ 关键变量
wm $\leftarrow$ WorldModel(检查点，使用 KV_缓存 =True)
wm.重置(x_0) ▷ 清除 KV缓存
对每个动作 a_t **执行**
wm.清除_kv_缓存() ▷ 每块缓存
将带噪潜变量追加到历史记录
开始 $\leftarrow$ max(0, 当前帧 +块 −最大帧数)
对每个扩散步**执行**
缓存 idx $\leftarrow$干净帧边界（开始） ▷ t=0 窗口
$v_{\text{cond}} \leftarrow$ DiT(x, t, a_t, 缓存 idx, 开始, 缓存 ="cond")
$v_{\text{null}} \leftarrow$ DiT($x, t,$ null(a_t), 缓存 idx, 开始, 缓存 ="null")
$v \leftarrow$ 分类器自由引导(v_{cond}, v_{null})
$x \leftarrow$ DDIM 更新(x, v，也许缓存后缀)
循环结束
解码最新的干净帧
当前帧 $\leftarrow$ 当前帧 +块
循环结束

A. 模型与训练的补充详情

A.1. 世界模型的快速推理

强化学习训练流程中的一个主要瓶颈是世界模型的推理延迟。未经优化时，时序注意力层会在每一步都对整个帧历史重新计算注意力。我们通过引入键值（KV）缓存（Quevedo 等人，2025年），修改了 WorldGym（Pope 等人，2023年）架构，以实现快速推理，具体如算法 1所述。这一改动显著降低了每步延迟，尤其是在长上下文情况下，使原本无法实现的长时域强化学习训练成为可能。实际应用中，我们观察到 10× 在 40 帧上并行生成 20 条轨迹时，rollout 时间有所缩短。

A.2. OpenVLA-OFT 作为基础模型

在所有实验中，我们使用 OpenVLA-OFT（Kim 等人，2025年）作为基础模型。OpenVLA-OFT 在 OpenVLA（Kim 等人，2024年）的基础上，提供了一套旨在提升训练效率和性能的微调方案。具体而言，它支持并行解码、动作分块和连续动作表示，从而提升推理和学习的速度与性能。我们从在 Open X-Embodiment 数据集（Vuong 等人，2023年）上预训练的 OpenVLA 检查点出发，使用 Bridge Dataset V2（Walke 等人，2023年）对 OpenVLA-OFT 进行微调。

A.3. 视觉语言模型作为奖励的细节

A.3.1. 用于奖励的视觉语言模型提示词

ˆ将 GPT-4o（Hurst 等人，2024年）用作奖励 R。

以下是机器人策略在基于视频生成的世界模型中展开运行时的一系列帧。我需要你帮助判断该策略是否成功。机器人完成以下任务的情况如何？指令：{instruction}{rubric.strip()}请简要说明理由（2--3 句话）。然后严格输出一行最终结果：最终得分：X，其中 X 是 ｛ ’0 或 1’ ｝之一。该行之后不要再输出其他数字。

注意：由于该视频由视频预测模型（以机器人动作为条件）生成，因此可能包含一些由视频模型容量造成的伪影。

评分标准如下：

评分标准：`0` =失败：未完成指令“{instruction}”。`1` =成功：已完成指令。

对于每次策略评估，我们在世界模型中执行 5 次相互独立的 rollout。为减少令牌使用量并缓解相邻帧中的冗余，每个 rollout 视频在发送给视觉语言模型进行评分之前，都会以步长 3 进行时间下采样。

A.4. 世界模型与强化学习训练的超参数详情

世界模型实现细节。WorldGym（Quevedo 等人，2025年）在我们的所有实验中均作为基于世界模型的模拟器。它使用来自 Stable Diffusion 3 的预训练变分自编码器（256×256 Esser 等人，2024年）将 RGB 图像帧编码到潜在空间中。WorldGym 的底层世界模型是一个 16 层 Transformer，隐藏维度为 1024，包含 16 个注意力头。除非另有说明，我们均采用默认的 WorldGym 配置进行视觉编码、世界模型架构和展开设置。我们使用基于 Open-X Embodiment 数据集训练的预训练检查点初始化世界模型（Vuong 等人，2023年）。

WorldGym 将动作表示为 10 维向量。在我们的实验中，我们使用前 7 个维度，以匹配 OpenVLA（Kim 等人，2024年）的动作空间，该空间由 6 维末端执行器位姿和二元夹爪状态组成。WorldGym 使用固定的 20 帧上下文长度进行训练；在较长的 rollout 期间，它以最近 20 帧的滑动窗口作为条件。

超参数	值
参数总量	609 M
图像分辨率	256×256
DiT 补丁大小	2
输入通道	16
隐藏维度	1024
层数	16
注意力头	16
MLP 比率	4
优化器	AdamW（权重衰减 $= 0.002, \beta_1 = 0.9, \quad \beta_2 = 0.99$）
学习率	8e-5
批大小	16
动作维度	10
训练硬件	2xA100 80GB
训练步数	580k
扩散噪声调度	Sigmoid函数
采样时间步	10
预测目标	v

表4. 用于训练 World-Gymnast 视频世界模型的超参数。

强化学习训练实现细节。我们在 WorldGym（Shao 等人，2024年）中使用组相对策略优化（GRPO）（Quevedo 等人，2025年）训练策略。对于每项任务，训练均从单个观测帧和一条语言指令开始。策略在世界模型中最多 rollout 40 步，并使用 GPT-4o（Hurst 等人，2024年）为每次 rollout 赋予二元任务完成奖励。动作以长度为 5 的块生成，策略更新采用裁剪策略梯度目标。所有强化学习训练均完全在 WorldGym 中进行。

超参数	值
强化学习算法	GRPO
世界模型环境	WorldGym
基础策略	OpenVLA-OFT
最大轨迹展开长度	40 步
动作块长度	5
奖励类型	二元任务完成
奖励来源	GPT-4o
组大小	8
批大小	20
学习率	5×10^{-6}
裁剪比率	$\epsilon_{\text{high}} = 0.28,\ \epsilon_{\text{low}} = 0.2$
温度	1.6
熵系数	0.0
梯度裁剪	1.0
优化器	AdamW
训练硬件	4 ×NVIDIA H200（140GB）
训练时长	1–2 天

*T*表 *5.* WorldGym 中强化学习训练的超参数 nast。

B. 基线

B.1. 监督微调

由于样本效率较高，监督微调是强化学习之外的一种策略改进方法。作为第一个基线，我们遵循 OpenVLA-OFT（Kim 等人，2025年）的方案，使用 Bridge V2 数据集（Kim 等人，2024年）中的专家轨迹对 OpenVLA-7B 模型进行微调（Walke 等人，2023年）。

诸如 Ctrl-World（Guo 等人，2025年）的方法探索了使用由世界模型生成的合成数据进行迭代式监督微调。遵循他们的方法，我们选取强化学习训练中使用的相同留出场景，并使用在 OpenX Embodiment 数据集（Vuong 等人，2023年）上预训练的世界模型，让基础监督微调策略在其中展开。为确保公平比较，迭代监督微调的展开步数与强化学习微调期间的总展开步数相同。随后，我们提示视觉语言模型筛选出成功轨迹，并将这些轨迹与 Bridge V2 数据集混合，用于下一轮微调，采样比例为 $1:1$。

B.2. SIMPLER

为了将 World-Gymnast 与基于模拟器的强化学习进行比较，我们选择了 SIMPLER 模拟器（Li 等人，2024年），它提供了 Bridge 机器人配置的数字孪生。SIMPLER 包含以下任务：1）将勺子放在桌布上，2）将胡萝卜放在盘子上，3）将绿色方块叠放在黄色方块上，4）将茄子放入篮子，5）将茄子放入水槽，6）打开抽屉和 7）关闭抽屉。最后 4 项任务是 AutoEval 配置的数字孪生（Zhou 等人，2025年）。

基础监督微调策略在 SIMPLER 任务上的表现不佳，导致奖励方差较低，且使用 GRPO 进行训练时不稳定（Shao 等人，2024年）。为解决这一问题，我们将轨迹奖励定义为步骤奖励之和。我们采用稠密奖励机制为每一步赋予部分奖励：满足以下各项条件时，分别奖励 0.1：是否抓取了源物体，是否已抓取，连续抓取，是否将物体明显抬起以及抬起物体。成功则奖励 1。

C. 数据集

C.1. OpenVLA 评测任务集

我们在一组精心整理的桌面操作任务上评估了不同方法，这些任务最初由 OpenVLA 提出（Kim 等人，2024年）。该任务套件旨在从多个维度评估策略泛化能力，包括视觉、运动、物理和语义变化，以及在包含多个物体的场景中的语言落地。对于强化学习训练，我们将该任务套件按 80/20 划分为训练集和评估集；两个子集包含相同的任务，但初始帧不同。

任务	泛化类型
将茄子放入锅中（简单版）	视觉泛化
将茄子放入锅中	视觉泛化
将杯子从台面放入水槽	视觉泛化
将茄子放入锅中（含杂物）	视觉泛化
将黄玉米放在粉色盘子上	视觉泛化
提起茄子	运动泛化
将胡萝卜放在盘子上（高度变化）	运动泛化
将胡萝卜放到盘子上	物理泛化
将锅翻正	物理泛化
拿起 AAA 电池	物理泛化
将头骨移入晾晒架	语义泛化
拿起白色胶带	语义泛化
将紫葡萄从锅中取出	语义泛化
将蓝色杯子叠放在粉色杯子上	语义泛化
将茄子、红色瓶子放入锅中	语言落地
提起奶酪、红辣椒	语言落地
将蓝色杯子、粉色杯子放到盘子上	语言落地

表6. OpenVLA 评测任务集的任务及相应泛化类型。

C.2. 视觉干扰物

为评估对视觉扰动的鲁棒性，我们通过修改来自 OpenVLA 评测任务集（Kim 等人，2024年）的初始帧，构建了干扰物增强训练分布。借助 Nano Banana（Google，2025年），我们在场景中插入视觉上多样但与任务无关的物体，同时保持底层任务、目标规范和机器人配置不变。干扰物数据集由 OpenVLA 评测任务集的训练集划分构建而成。具体而言，90% 的训练回合使用原始初始帧，而 10% 则从干扰物增强帧进行初始化。

在评估中，我们同时在一组加入干扰物的留出集初始帧上，以及在 WorldGym 中原始的 OpenVLA评估任务上（Quevedo 等人，2025年）测量性能，以评估视觉扰动下的鲁棒性。图 2 中展示了视觉干扰物下的定性 rollout 示例

用于编辑初始帧的 Nano Banana 提示词。

这是机器人任务 [TASKINSTRUCTION] 的初始观测。请以自然的方式在场景中添加干扰物来修改此图像，但不要移动或改变原始场景中的任何物体。要求：

- 机械臂应清晰可见。
- 即使存在干扰物，任务 [TASK INSTRUCTION] 仍应能够完成。

C.3. 分布外语言

为了评估对语言分布偏移的鲁棒性，我们通过修改部分 OpenVLA 任务的任务指令，构建了一个分布外（OOD）语言训练分布（Kim 等人，2024年）。底层场景和物体保持不变；只有自然语言指令经过修改，以描述新的交互。

分布外语言数据集由四种新定义的语言变体组成，应用于现有的 OpenVLA 任务：**(a)**：把装有葡萄的锅移到晾晒架上，**(b)**：从晾晒架上拿起锅，放到水槽外的台面上，**(c)**：把盘子放到晾晒架上，以及**(d)**：把黄玉米放进红杯子里。在强化学习训练期间，我们将此分布外语言数据与 OpenVLA 评测任务集的训练集划分结合，在原始语言指令和分布外语言指令之间，按初始帧进行 50/50 混合。图 6中展示了分布外语言指令下的定性 rollout 示例。

C.4. 评估 World-Gymnast 提供的多样化设置

C.5. 扩展任务集

为研究在强化学习训练期间增加任务多样性的影响，我们通过在 OpenVLA 评测任务集的训练集划分中加入从 Bridge Dataset V2 随机选取的五项额外操作任务，构建了一个扩展训练分布（Walke 等人，2023年）：**(a)**：关上柜门，**(b)**：把橙色的锅翻转过来，正放在水槽里，**(c)**：从右向左折叠布，**(d)**：打开抽屉，以及**(e)**：把水槽盘子上的锅铲拿开。这些额外任务不属于 OpenVLA 评测任务套件，仅在强化学习训练期间使用，而训练流程保持不变。

C.6. 真实机器人评估数据集

为评估真实世界迁移，我们报告了在 AutoEval（Zhou 等人，2025年）上的性能。AutoEval 是一个自动化的真实机器人评估框架，其任务取自 Bridge Dataset V2 分布。在实验中，我们使用 WidowX机械臂，在两个场景中评估 AutoEval 支持的三项操作任务：两项抽屉操作任务（打开抽屉，关闭抽屉）以及水槽场景中的一项抓取放置任务（将茄子放入蓝色水槽）。我们未对 AutoEval 支持的布料操作任务进行评估。扩展任务集上的定性 rollout 示例见图5。

D. 评估

D.1. WorldGym 评估

我们在 WorldGym（Quevedo 等人，2025年）中评估学习得到的策略。WorldGym 是一种基于世界模型的模拟器，旨在近似真实机器人执行。

展开协议。 对于每项任务，策略都会从给定的初始帧和语言指令出发，在 WorldGym 中展开，最多执行 40 步。每次展开都会生成一系列预测的未来帧，用于评估任务是否完成。

成功评估。 按照 WorldGym 评估协议，使用视觉语言模型（VLM）将 rollout 的结果与任务指令进行比较，以判定任务是否成功。为降低基于 VLM 的评分方差，我们对每次 rollout 获取 5 个独立的 VLM 判断，并通过多数表决赋予二元成功标签。

D.2. AutoEval 评估

AutoEval（Zhou 等人，2025年）提供了一个便捷可靠的平台，可在真实世界机器人上进行策略评估。平台提供两套设备，每套都配备一台 WidowX 机械臂（图 5）。每套设备都可执行两个操作相反的任务（例如，打开和关闭抽屉）。在评估某个任务的策略时，可以使用重置策略重置条件，以便被评估的策略进行下一次试验。

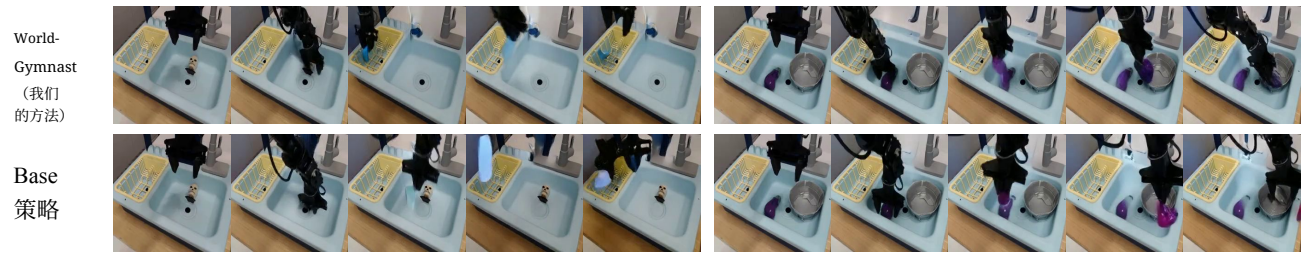


图 *4.* **WorldGym 中策略 rollout 的定性评估。** 我们比较了经过强化学习微调的 World-Gymnast 策略与执行强化学习前的基础策略。**左：** 抬起头骨；**右：** 将茄子放入锅中。

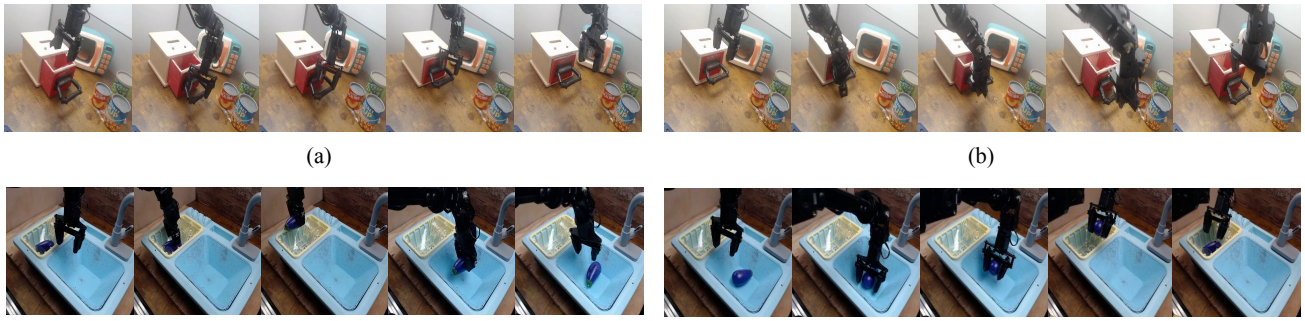


图 *5.* **AutoEval 中策略 rollout 的定性评估。(a)：** 关闭抽屉；**(b)：** 打开抽屉；**(c)：** 将茄子放入蓝色水槽；**(d)：** 将茄子放入黄色篮子。

E. 不同场景下的定性示例

本节将提供更多定性结果和可视化内容，以补充我们报告的结果。

轨迹可视化。在图4中，我们展示了在学习到的世界模型 WorldGym 中的一些轨迹样本。我们还对比了基础策略与 World-Gymnast 的强化学习微调策略，说明了基于世界模型进行强化学习的有效性。图5 展示了一些 AutoEval 评估设置的示例。另一方面，图9 展示了 WorldGym 在略微分布外的 AutoEval 图像帧上的表现。

World-Gymnast 的多样化训练场景。我们在图6中提供了关于前文讨论的、由 World-Gymnast 支持的多样化训练设置的更多定性结果。图7展示了在世界模型中未见过的场景上进行训练且不使用专家数据时，策略的行为。

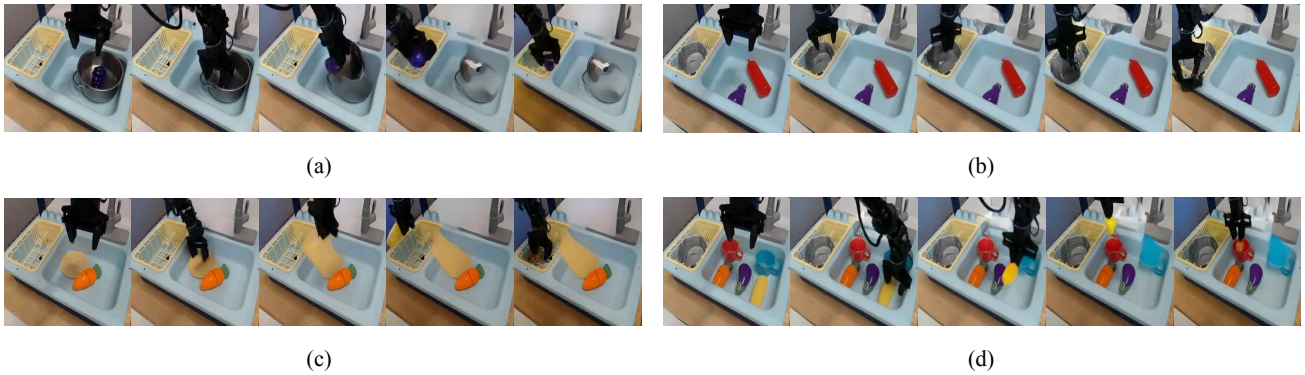
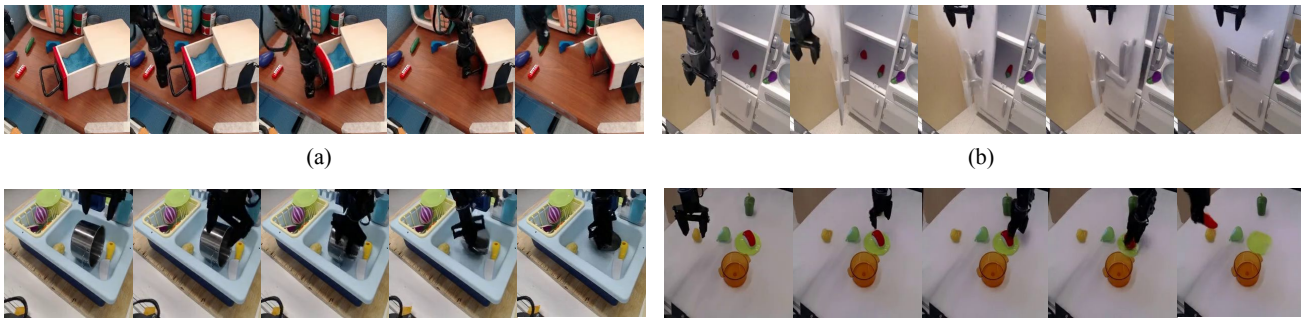


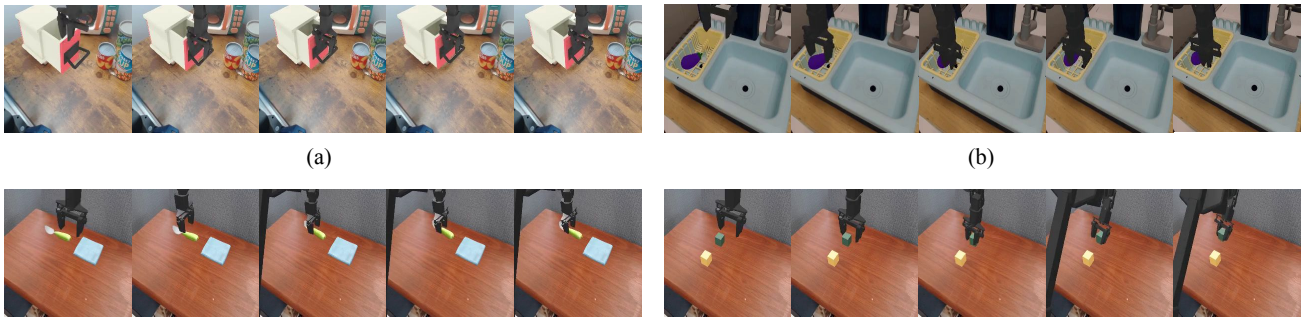
图 *6.* **使用分布外语言描述对策略 rollout 进行定性评估。(a)：** 将装有葡萄的锅移到晾晒架上；**(b)：** 从晾晒架上拿起锅，放到水槽外的台面上；**(c)：** 将盘子放到晾晒架上；**(d)：** 把黄玉米放进红色杯子里。



(a) (b)

(c) (d)

图7. **在新颖的初始图像帧下，对策略 rollout 进行定性评估。(a)：**关闭抽屉；**(b)：**关闭冰箱；**(c)：**将水槽里的锅翻转至正立；**(d)：**把盘子上的寿司拿走。



(a) (b)

(c) (d)

图 8. **在 SIMPLER 模拟器中训练的策略的定性评估。(a)：**关闭抽屉；**(b)：**把茄子放进水槽；**(c)：**把勺子放在桌布上；**(d)：**将绿色方块叠放在黄色方块上。



(a) (b)

图9. **使用 AutoEval 的帧对 WorldGym 中的策略 rollout 进行定性评估。(a)：**把茄子放进蓝色水槽；**(b)：**打开抽屉。WorldGym 未使用这些来自 AutoEval 的图像观测进行训练。

与 SIMPLER 的比较。这里我们在图8 中展示了更多定性评估，评估的是在 SIMPLER 模拟器中使用强化学习训练的策略，具体讨论见第4.2节。在此类模拟器中进行训练，不太适合实际部署。