计算机科学
COMPUTER SCIENCE

基于多模态大模型辅助视频动作生成的预训练世界模型

万盛华, 徐兴业, 甘乐, 詹德川

引用本文

万盛华, 徐兴业, 甘乐, 詹德川. 基于多模态大模型辅助视频动作生成的预训练世界模型[J]. 计算机科学, 2026, 53(1): 51-57.

WAN Shenghua, XU Xingye, GAN Le, ZHAN Dechuan. Pre-training World Models from Videos with Generated Actions by Multi-modal Large Models [J]. Computer Science, 2026, 53(1): 51-57.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250800033

基于多模态大模型辅助视频动作生成的预训练世界模型

万盛华　徐兴业　甘　乐　詹德川
南京大学人工智能学院　南京 210023
南京大学计算机软件新技术国家重点实验室　南京 210023
(wansh@lamda.nju.edu.cn)

摘　要　预训练世界模型是提升强化学习样本效率的关键技术,但现有方法因视频数据缺乏显式动作标注,难以捕捉状态转移的因果机制。对此,提出多模态大模型辅助的视频动作生成预训练框架(MLM-generated Action-based Pre-training from videos for world models,MAPO),通过整合视觉语言模型的语义理解能力与动力学建模需求,突破传统预训练范式在动作语义缺失方面的局限性。具体地,MAPO在预训练阶段利用多模态大模型(QWEN2_5-VL-7B)解析视频帧序列,生成细粒度语义动作描述,构建具有因果解释性的动作-状态关联;设计上下文量化编码机制,解耦场景静态特征与动态控制因素,增强跨模态表征能力。在微调阶段,通过双网络协同架构实现预训练动力学特征与真实环境动作的端到端对齐。实验表明,MAPO在DeepMind Control Suite和Meta-World的8项任务中的平均回报较最优基线获得稳定提升,尤其在长时程任务中展现出卓越的性能。该研究为跨模态世界模型训练提供了新范式,揭示了语义动作生成在因果推理中的关键作用。

关键词:世界模型;强化学习;视频预训练;多模态大模型;语义动作生成
中图分类号　TP183

Pre-training World Models from Videos with Generated Actions by Multi-modal Large Models

WAN Shenghua,XU Xingye,GAN Le and ZHAN Dechuan
School of Artifcial Intelligence,Nanjing University,Nanjing 210023,China
State Key Laboratory for Novel Software Technology,Nanjing University,Nanjing 210023,China

Abstract　Pre-training of world models is key to improving the sample efficiency of reinforcement learning. However,existing methods struggle to capture the causal mechanisms of state transitions due to the lack of explicit action labels in video data. This paper presents MAPO(Multimodal-large-model-generated Action-based pre-training from videOs for world models),a novel pre-training framework. It leverages the semantic understanding of visual-language models and meets the needs of kinematic modeling,overcoming the limitations of traditional pre-training methods in the absence of action semantics. MAPO uses the multimodal large model(QWEN2_5-VL-7B) to analyze video frame sequences and generate fine-grained semantic action descriptions during pre-training. This establishes action-state associations with causal explanations. It also designs a context quantization encoding mechanism to separate static scene features from dynamic control factors,improving cross-modal representation. During fine-tuning,MAPO uses a dual-network collaborative architecture to align the pre-trained kinematic features with real-environment actions. Experiments show MAPO steadily improves average returns over baselines in 8 tasks on DeepMind Control Suite and Meta-World,especially in long-horizon tasks. This study offers a new cross-modal world model training approach,highlighting the importance of semantic action generation in causal reasoning.

Keywords　World models,Reinforcement learning,Video pre-training,Multi-modal large models,Semantic action generation

1　引言

在基于模型的强化学习(Model-Based Reinforcement Learning,MBRL)框架中[1-2],世界模型[3]作为智能体对物理环境进行认知推理的核心组件,通过构建可预测的环境动力学表征,为决策策略提供高效的想象训练空间[4-5]。这种基于世界模型的规划范式,能够显著减少与真实环境的试错交互次数,成为提升样本效率的关键技术路径[5]。近年来,随着自动驾驶[6]和机器人操作[7]等复杂决策任务需求的增长,构建能够精准预测多步状态转移的通用世界模型已成为人工智能

到稿日期:2025-08-11　返修日期:2025-10-20
基金项目:国家自然科学基金青年学生基础研究项目(博士研究生)(624B200197);国家重点研发计划(2022ZD0114805)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(Ph.D Candidate)(624B200197) and National Key Research and Development Program of China(2022ZD0114805).
通信作者:詹德川(zhandc@nju.edu.cn)

领域的重要研究方向[8]。

当前，世界模型的训练高度依赖海量的环境交互数据[9]。然而，真实场景数据的采集面临成本高昂、安全风险大等现实约束[10]。这促使研究者转向利用仿真环境生成合成数据，通过高保真物理引擎构建的虚拟场景，能够快速获取大量训练样本[11]。但此类方法存在显著的模型迁移鸿沟：基于仿真数据训练的世界模型在参数化建模过程中不可避免地会继承虚拟环境的特定假设，难以捕捉真实世界的复杂动力学特征[12]。为突破这一限制，跨模态迁移学习成为新的研究热点——通过整合多源异构数据（如不同传感器数据、跨场景操作记录等）训练具有泛化能力的统一世界模型[13]。然而，由于不同模态数据在采集方式、状态表征和动作空间等方面存在显著差异[14]，现有方法在跨模态知识融合过程中普遍面临表征对齐困难、潜在空间解耦不足等挑战。

近期，研究者开始探索可扩展的通用世界模型构建方法。TD-MPC2 作为一种隐式（Decoder-free）世界模型方法，通过在潜在空间执行局部轨迹优化，成功实现了在 104 个多样化任务上的稳健表现，并提出了“预训练-微调”的范式来提升多任务世界模型的能力[8]。该工作引入了可学习任务嵌入空间和动作掩码机制，以处理多个观察和动作空间对齐的问题，展示了模型与数据规模扩大带来的性能提升。然而，TD-MPC2 仍然依赖环境交互数据进行预训练，难以充分利用海量无标注视频数据中蕴含的丰富视觉动态信息。与此同时，GAIA-1 将世界建模任务简化为类似大型语言模型的“预测下一个 token”任务，通过向量量化表示将视频帧离散化为 token 序列，已成功应用于自动驾驶场景[6]。GAIA-1 展示了世界模型的可扩展性，证明了类似 LLM 的缩放规律同样适用于复杂环境建模。但这种方法将世界模型简化为序列预测任务，割裂了视频数据与决策任务在动作语义层面的关联，难以建立真实的行为-状态映射关系，特别是在需要精确动作控制的场景中表现受限。

针对上述问题，近期研究提出“预训练-微调”的创新范式：先在海量视频数据上构建通用世界模型，再通过少量领域数据微调实现下游任务适配[15-16]。代表性工作 APV[15] 通过无监督学习从无标注视频中提取潜在动力学特征，但其在微调阶段需要重新构建动作条件模型，导致知识迁移效率低下。PreLAR[16] 尝试通过伪动作生成弥补动作信息缺失，但视频帧间的伪动作关联易产生时序因果关系混淆（Temporal Confounding），难以建立真实的行为-状态映射关系。这些局限性本质上源于视频数据与决策任务在动作语义层面的割裂——传统视频预训练仅关注视觉动态变化，而忽略了驱动状态转移的因果动作因素[17]。

本文提出的多模态大模型辅助的视频动作生成预训练框架（MAPO），通过深度整合视觉语言模型的语义理解能力与强化学习的动力学建模需求，开创性地解决了跨模态预训练中的动作语义缺失问题。具体而言，本方法在预训练阶段利用多模态大模型对视频帧序列进行细粒度动作语义解析，构建具有因果解释性的动作-状态关联；在微调阶段则通过自适应特征融合机制，实现预训练动力学表征与真实环境动作空间的平滑衔接。这种双阶段协同训练范式，不仅解除了传统方法对显式动作标注的依赖，更重要的是建立起了从视频语义到控制指令的端到端可微映射，为世界模型赋予了真实物理场景中的因果推理能力。

本文的主要贡献包括 3 个方面：

1）提出基于语义动作生成的跨模态预训练新范式，首次实现了无标注视频数据与决策任务的深度语义的对齐；

2）设计多尺度上下文量化编码机制，有效解耦场景静态特征与动态控制因素，增强了模型对复杂环境的表征能力；

3）所提框架在多个标准测试环境中取得稳定的性能提升，特别是在长时程任务中展现出优异的性能。

2 相关工作

2.1 基于模型的强化学习

MBRL 通过构建环境动力学模型来提升决策能力，减少与真实环境的试错交互，提高学习效率。世界模型通过近似实际环境的状态转移和奖励预测来增强动力学模型，利用想象的轨迹促进规划和行为学习[4-5]。Ha 和 Schmidhuber 首次提出将世界模型[3]定义为紧凑表示空间内的潜在动力学模型，其通过变分自编码器[17]从视觉观察中压缩得到潜在状态表示。后续研究，如 Dreamer 系列[4-5]，通过引入随机动力学、离散潜在表示和对数似然预测来增强这一模型，使其在各种领域中稳定应用。最近的研究[18-21]将 Transformer 架构[22]纳入世界模型，结合其强大的序列建模和生成能力，在 Atari 100k 基准测试[23]中实现了更高性能。此外，状态空间模型也被用于提高世界模型的长期序列建模能力[24]。一些研究还提出将世界模型设计为动作条件视频预测模型，基于先进的扩散模型[25]生成想象轨迹[26-27]。尽管如此，现有的世界模型研究仍需要与实际环境进行大量交互，其效率受到限制。本文充分发挥了预训练阶段的潜力来增强世界模型的性能。

2.2 预训练世界模型

为了提高世界模型的学习效率，先预训练再微调的流程提供了可行的解决方案。APV[15]首次研究了世界模型的预训练和微调范式：在预训练阶段，从无动作视频中训练一个无动作动力学模型；在微调阶段，叠加一个基于动作的动力学模型在无动作模型之上，并对整个模型进行联合优化，以适应下游任务。ContextWM[28]进一步改进了这一框架，在预训练期间分别对视觉上下文和动力学进行单独建模，从复杂多样的视频中学习。SWIM[29]在大规模人类操作数据集上预训练动作条件世界模型，并将其迁移到机器人操作任务中，使用现成的视觉模型从操作视频中显式检测动作标签。iVideoGPT[18]采用 Transformer 架构，将视频转为视觉令牌并进行并行处理，实现高效长序列建模；通过自回归预测与动作条件输入，兼顾可扩展性与交互性，支持长时规划与实时响应；模型在百万级操作轨迹上训练，适用于复杂操控任务。但离散令牌可能损失细节，且因果关系依赖数据隐式关联，存在建模偏差风险。PreLAR[16]则提出了一个动作条件预训练方案，通过从无动作视频中学习动作表示，弥合了预训练和微调之间的差距，从而更有效地将预训练模型的知识转移到特定的下游任务中。这些方法通过预训练增强了世界模型的表示能力，提高了学习效率。然而，预训练视频缺少动作，很难让世界模型真正捕捉到时序上观测产生变化的内在因果关系，即便是学习伪动作的 PreLAR 方法，也存在视频帧之间的捷径关系

(shortcut)，并非是真实的动作原因。本文延用了预训练再微调的范式，在两个阶段均基于多模态大模型辅助生成多帧视频上的语义动作，帮助世界模型捕捉帧之间的因果关系。

3 预备知识

3.1 强化学习

一个标准的强化学习问题可以定义为马尔可夫决策过程，即一个元组 $M=(S,A,P,p,r,\gamma)$，其中 S 是状态空间，A 是动作空间，$P:S\times A\to S$ 是转移函数，p 是初始状态分布，$r:S\times A\times S\to\mathbb{R}$ 是奖励函数，γ 是折扣因子。强化学习的目标是学习一个最优策略 $\pi(a|s)$，以最大化期望的累积折扣回报 $R_M(\pi)=E[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1})]$。为了解决不完全观测问题(如高维图像)，以往的研究采用了部分可观测马尔可夫决策过程(POMDPs)假设[30]。该假设引入了观测空间 O，并通过 $\phi:S\to O$ 生成观测。在本工作中，由于模型接收来自环境的图像观测，因此采用 POMDPs 假设来建模环境。

3.2 世界模型

视觉模型强化学习(MBRL)通过构建环境动态模型(即世界模型)提升样本效率[4-5]。其核心是学习一个潜在动态模型(Latent Dynamics Model)，用于预测状态转移和观测生成。形式化定义如下。

给定观测序列 $\{o_t\}_{t=1}^{T}$ 和动作序列 $\{a_t\}_{t=1}^{T}$，世界模型由以下组件构成。

1)表示模型：$z_t\sim q_\theta(z_t\mid z_{t-1},a_{t-1},o_t)$，从历史状态和当前观测推断潜在状态。

2)转移模型：$\hat{z}_t\sim p_\theta(\hat{z}_t\mid z_{t-1},a_{t-1})$，预测下一潜在状态。

3)图像解码器：$\hat{o}_t\sim p_\theta(\hat{o}_t\mid z_t)$，从潜在状态重建观测。

4)奖励预测器：$\hat{r}_t\sim p_\theta(\hat{r}_t\mid z_t)$，预测即时奖励。

模型通过优化变分下界联合训练：

$$\mathcal{L}(\theta)=\mathbb{E}_{q_\theta}[\sum_{t=1}^{T}(-\ln p_\theta(o_t\mid z_t)-\ln p_\theta(r_t\mid z_t)+\beta_z KL[q_\theta\parallel p_\theta])] \tag{1}$$

其中，β_z 为 KL 散度权重。

4 基于多模态大模型辅助视频动作生成的世界模型预训练与微调方法

本文提出基于多模态大模型辅助视频动作生成的世界模型预训练方法 MAPO，利用多模态大模型，在预训练阶段对视频数据集以及在下游任务微调阶段对多帧环境观测图像，分别生成动作向量编码，参与世界模型中任务无关的环境转移动力学网络的训练过程，提升世界模型对视频语义和物理规律的理解能力，实现高效知识迁移与样本高效学习。本章从模型架构、上下文量化编码以及训练流程 3 方面展开介绍。MAPO 的伪代码如算法 1 所示。

算法 1 MAPO

输入：视频数据集 D_{video}，下游任务环境 E，多模态大模型 MLM

输出：优化的策略网络 π_ψ

1. / * 预训练阶段 * /
2. 初始化：语义动作网络 f_φ，ResNet 编码器，解码器，大小为 M 的码簿 CB
3. for each 训练步：
4. 从 D_{video} 中采样视频片段 $o_{1:T}$
5. / * 上下文量化编码 * /
6. 随机采样单帧 o_τ，提取上下文 $c_\tau\leftarrow$ResNet(o_τ)
7. 查询码簿得到上下文编码 $e_i\leftarrow\arg\min_i\parallel c_\tau-e_i\parallel_2$
8. / * 基于语义动作训练世界模型 * /
9. for t=1 to T：
10. 生成语义动作 $\hat{a}_t\leftarrow$MLM($o_{t-4:t}$)
11. 通过最小化损失 $L_{pre\text{-}train}$ 更新 f_φ 参数
12. end for
13. 更新码簿 CB
14. end for
15. / * 微调阶段 * /
16. 初始化：环境动作网络 f_θ，演员网络 π_ψ，评论家网络 v_ξ
17. for each 训练步：
18. 从环境 E 收集真实轨迹 (o_t,a_t,r_t)
19. 联合优化 f_φ 和 f_θ 最小化损失 $L_{fine\text{-}tune}$
20. / * 策略学习阶段 * /
21. 基于优化后的世界模型(f_φ,f_θ) 生成潜在轨迹
22. 计算 λ-折现回报 V_τ^λ
23. 通过最小化 L_{critic} 更新评论家网络 v_ξ
24. 通过最大化回报(最小化 L_{actor})更新演员网络 π_ψ
25. end for
26. 返回 π_ψ

4.1 模型架构

MAPO 基于堆叠循环架构(见图 1)，包含以下核心模块。

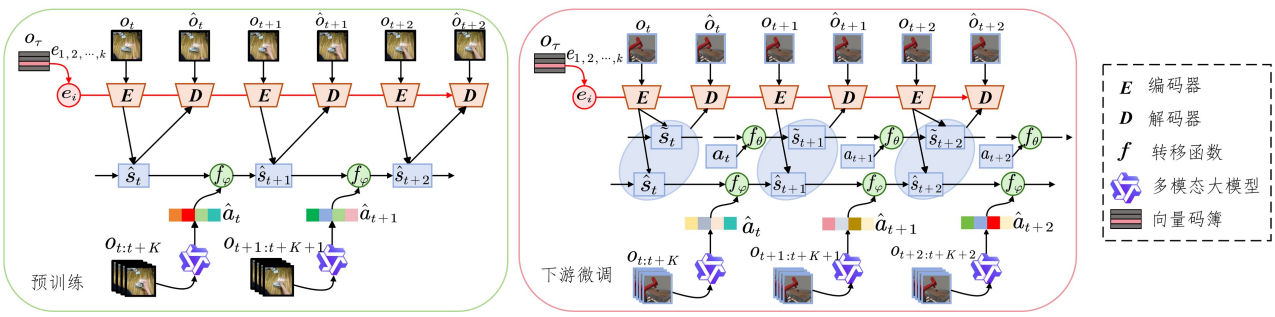


图 1 MAPO 方法的整体框架

Fig. 1 Overall framework of MAPO

1)语义动作预训练网络 f_φ：从互联网视频中学习潜在动态与上下文表征，建模潜在状态 s_t，通过变分推断学习 $q_\varphi(s_t\mid s_{t-1},o_t,\hat{a}_t)$ 与 $p_\varphi(\hat{s}_t\mid\hat{s}_{t-1},\hat{a}_t)$，其中 $\hat{a}_t$ 由 QWEN_VL 2_5 模型[31]基于连续 4 帧图像 $o_{t-4:t}$ 生成，与 sin-cos 时

序位置编码 pos_t 相加得到。

2)环境动作微调网络 f_θ:在微调阶段叠加动力学转移模型 $\tilde{s}_t \sim q_\varphi(\tilde{s}_t \mid \tilde{s}_{t-1}, a_{t-1}, \hat{s}_t)$,在下游任务中引入真实的环境动作,生成可控轨迹。该设计支持预训练-微调范式:预训练阶段仅优化视频语义动作网络层;微调阶段联合优化两个网络的参数,保留预训练的视觉编码器与解码器权重。

4.2 上下文量化编码

视频通常多包含复杂静态背景(如物体纹理、光照条件)与动态变化(如视角切换)[28],这些信息可能和主体内容信息(运动控制)不相关。为有效捕捉视频的场景信息并区分不同场景,在 ContextWM 上下文编码[28]的基础上,MAPO 引入上下文量化编码 e_i,其建模方式如下:从视频片段 $o_{1:T}$ 中随机采样单帧 o_τ,$\tau \sim$ Uniform$(1, T)$,基于非预训练的上下文 ResNet 编码器通过 L 层残差块和下采样操作生成多尺度特征表示 c_τ 作为上下文[28]。上下文编码器在平均池化前的最后一个残差块输出(尺寸分别为 16×16 和 8×8)被传递到图像解码器的相应残差块。这些不同分辨率的特征被存储在 shortcuts 字典中,以空间维度为键,实现了对上下文信息的多粒度捕获。然后,MAPO 学习一个大小为 M 的码簿 CB,根据 VQ-VAE[32]的标准更新方式从中得到编码 e_i,$i \in \arg\min_i \| c_\tau - e_i \|_2$,$i \in \{1, \cdots, M\}$。在上下文量化编码的学习过程中,本文遵循 ContextWM 的假设[28],即每帧观测均包含完整的上下文信息,基于随机采样来增强模型的鲁棒性。本文的上下文量化编码具有以下两方面优势:1)上下文编码可视为针对视频的冗余场景信息的条件表示,以此减轻解码器图像重构的负担;2)基于 VQ 的量化编码可以有效统一相似度较高的视频表示,进一步增强编码的抽象能力。

4.3 模型架构

1)预训练阶段

仅在视频数据集上结合多模态大模型生成的语义动作 $\hat{a}$ 优化世界模型的语义动作网络,目标函数为:

$$\mathcal{L}_{\text{pre-train}} = \mathbb{E}_{q_\varphi}\left[\sum_{t=1}^{T}(-\ln p_\varphi(o_t \mid \hat{s}_t, e_i) + KL[q_\varphi \| p_\varphi])\right] \tag{2}$$

2)微调阶段

联合优化语义动作网络与环境动作网络,目标函数扩展为:

$$\mathcal{L}_{\text{fine-tune}} = \mathbb{E}_{q_\theta, q_\varphi}\left[\sum_{t=1}^{T}(-\ln p_\theta(o_t \mid \hat{s}_t, \tilde{s}_t, e_i) - \beta_r \ln p_\theta(r_t \mid \hat{s}_t, \tilde{s}_t) + KL[q_\varphi \| p_\varphi] + KL[q_\theta \| p_\theta])\right] \tag{3}$$

3)策略学习阶段

基于 DreamerV2[5]的演员-评论家框架,在模型想象出的潜在轨迹上优化策略。

(1)评论家。最小化价值函数误差:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}\left[\sum_{\tau=t}^{t+H}\frac{1}{2}(v_\xi(s_\tau) - sg(V_\tau^\lambda))^2\right] \tag{4}$$

其中,V_τ^λ 为 λ-折现目标。

(2)演员。最大化折现回报与熵正则项:

$$\mathcal{L}_{\text{actor}} = \mathbb{E}\left[\sum_{\tau=t}^{t+H}(-V_\tau^\lambda - \eta H[\pi_\psi(a_\tau \mid s_\tau)])\right] \tag{5}$$

5 实验与结果分析

本文在运动和抓取这两个场景共 8 个视觉强化学习任务上设计了实验。DeepMind Control Suite(DMC)是一个被广泛使用的机器人运动基准[33],本文选择了其中 4 个任务即 Cheetah Run, Walker Run, Hopper Hop 和 Quadruped Run 进行测试;Meta-world 是 50 个不同机器人操作任务的基准[34],本文使用了 Door Open, Drawer Close, Lever Pull 和 Dial Turn 这 4 个任务。本文旨在回答以下 4 个问题:1)MAPO 方法是否可以获得更好的任务性能? 2)MAPO 的各个设计对于提升性能是否有效? 3)多模态大模型输出的动作是否符合任务观测的变化? 4)上下文量化编码是否能得到高区分度的场景编码?

预训练数据集:本文选择了两个网络视频数据集来预训练世界模型,这些数据集可能对视觉控制有所帮助。其中,Something-Something-v2(SSv2)数据集包含 193000 个与物体互动的人类视频[35];Human3.6M 数据集包含 4 个不同视角下的人类姿势视频[36],共计超过 300 万帧。

对比方法:本文对比了 ContextWM[28], PreLAR[15], APV[16]和 iVideoGPT[18]这 4 种预训练方法以及标准的 DreamerV2[5]非预训练世界模型方法。

多模态大模型:本文选择了 QWEN2_5-VL-7B[31]。该模型是通义千问系列的多模态版本,具有 70 亿参数,支持多语言,能够处理文本、图像等多种输入形式,完成问答、创作、分析等任务。

动作提示词:本工作设计了提示词输出模板,用于生成格式化的动作描述(如"[手][打开][门][向前][10 厘米]")。原始提示词模板为:prompt = "Please describe possible actions(if there appears significant action then describe it else output null) in this video with format:[Executing body][Action][Target object][Moving direction][Displacement generated]. Example 1:A hand lifted a brick and moved it the distance of one arm. Example 2:A leg kicked the ball two meters away. Example 3:null."。这种提示词设置合理,覆盖了运动和操作任务中可能存在的动作行为模式信息,确保了多模态大模型生成的动作具有语义性和实际意义。

5.1 性能对比实验

为评估 MAPO 的性能,在 DMC 和 Meta-world 任务上将其与 ContextWM, PreLAR, APV 及 DreamerV2 进行了对比。图 2 所示结果表明,MAPO 在所有任务中均达到或超过对比方法的性能,长期任务稳定性优于基线方法,尤其在 Quadruped Run 这种较为复杂的运动任务中性能提升显著,MAPO 的平均回报较 ContextWM 高出约 7%,较 PreLAR(该方法学习逆动力学模型,基于前后帧推断隐动作)高出约一倍,且置信区间更窄。这得益于两方面:1)MAPO 通过多模态大模型生成的语义动作提升了世界模型的预测能力;2)MAPO 的量化编码可有效分离场景静态特征(如背景纹理)与动态控制因素(如物体运动),增强复杂环境的表征能力。这验证了 MA-

PO 结合预训练和语义动作生成的合理性，其能够有效提升下游任务的表现。

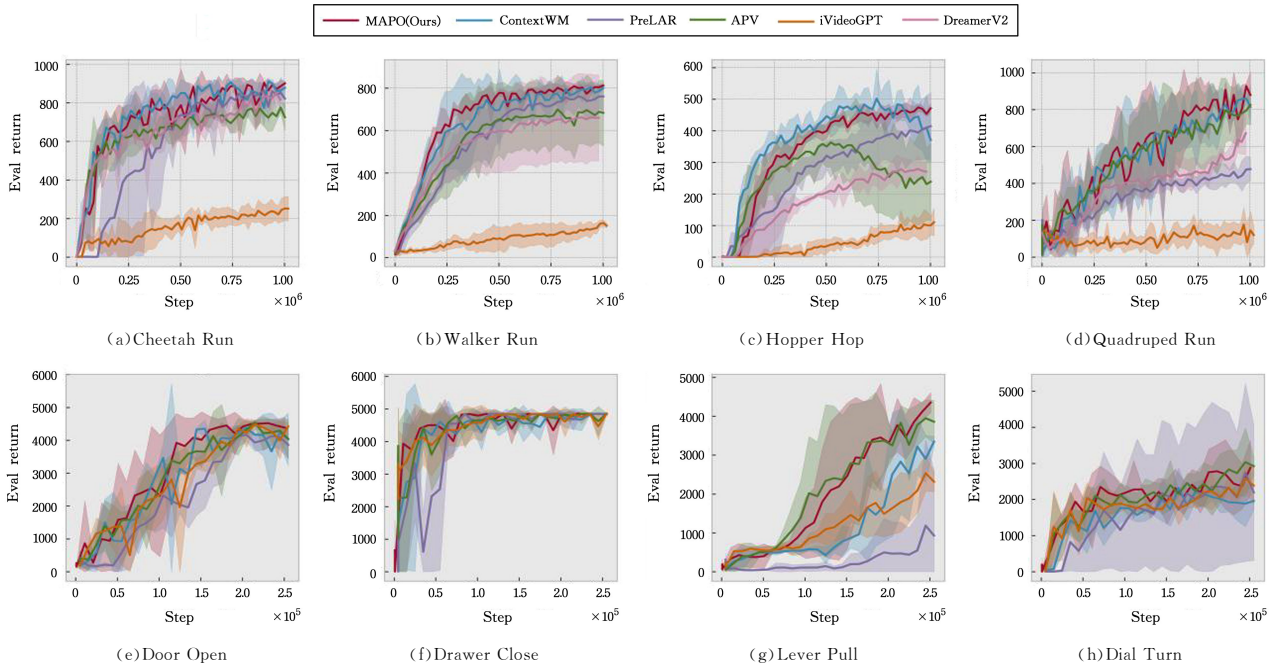


图 2 MAPO 与对比方法在多个任务上的性能表现(3 个随机种子的均值与 95%置信度区域)

Fig. 2 Performance of MAPO and baselines on several tasks(mean and 95% confidence interval with three random seeds)

基于 Transformer 架构的 iVideoGPT 方法[18] 尽管在 Meta-World 环境的部分抓取类任务上取得与 MAPO 相近的性能，但在多个 DMC 运动控制类任务上远差于 MAPO 方法。这是因为 iVideoGPT 预训练依赖无动作视频，仅学习视觉动态而忽略了动作驱动的因果机制；微调阶段虽引入动作输入，但线性投影的粗糙动作表示导致细粒度动力学预测效率低下，无法达成 DMC 的高精度控制需求。MAPO 方法通过多模态大模型生成结构化动作描述(如“[关节][弯曲][10 度]”)，显式建立动作-状态因果链，并通过上下文量化编码分离场景静态特征与动态控制因素，提升任务信息的表示能力。

5.2 消融实验

本节通过消融实验验证 MAPO 设计的有效性。最终测试的 10 条轨迹的平均性能如表 1 所列(为节约篇幅，任务的第一个单词仅保留首字母大写)。可以看出，去除上下文量化编码(CVQ)模块导致 MAPO 在 DeepMind Control Suite (DMC)的 4 个任务中性能显著下降，尤其以 Quadruped Run 最为明显(879.5→759.8)。这一现象印证了 CVQ 的核心作用：其通过 VQ-VAE[32] 的量化机制，将视频中的静态场景特征(如光照、背景纹理)与动态控制因素解耦，生成高区分度的上下文编码。

表 1 上下文量化编码的消融实验

Table 1 Ablation study on context quantized embedding

	CRun	WRun	HHop	QRun	平均
CVQ	**900.8**	**812.8**	471.0	**879.5**	**766.0**
w/o CVQ	856.7	794.9	**491.3**	759.8	725.7

在预训练数据集的影响分析中(见表 2)，SSv2 数据集在抓取类任务中的表现(Dial Turn：2414.8)优于运动类任务(Cheetah Run：866.9)；而 Human3.6M 则在运动类任务中表现更佳(Cheetah Run：900.8)。这一差异源于数据集的内在特性：SSv2 包含大量物体交互视频(如“开门”“旋转旋钮”)，其动作语义与 Meta-World 的抓取任务高度匹配；而 Human3.6M 聚焦人体骨骼运动，与 DMC 的肢体运动动力学结构更为契合。这验证了预训练数据与下游任务的语义对齐是提升迁移效果的关键——当数据集中蕴含的动作模式与目标任务的物理规律一致时，预训练模型能更有效地提取可泛化的动力学特征。

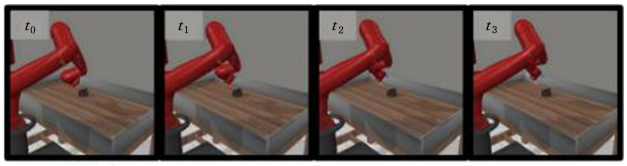
表 2 预训练视频数据集的影响

Table 2 Effects of pre-training video datasets

	MAPO		ContextWM	
	SSv2	Human3.6M	SSv2	Human3.6M
CRun	866.9	**900.8**	814.1	877.6
QRun	864.0	**879.5**	868.0	814.3
DTurn	**2414.8**	1958.0	2376.2	2338.3
DOpen	**4441.8**	4030.4	4382.7	4386.3

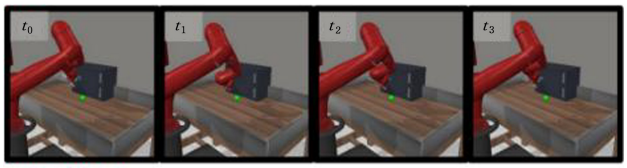
5.3 探索性实验

为检验生成的动作是否符合任务观测变化，实验随机抽取 Meta-World 任务中的连续多帧环境观测，由 QWEN 模型基于人为设计给定的提示词来输出动作描述。图 3 所示结果显示，在 Dial Turn 和 Door Open 任务中，QWEN 模型描述了开门动作的执行主体和位移方向。这得益于，多模态大模型在海量文本-视频数据集上进行了大规模的预训练[31]，基于视频帧能够生成细粒度语义动作描述(如“[手][打开][门]”)，建立视频帧序列与动作语义的因果关联。因此，多模态大模型能有效提取视频中的语义动作，为世界模型提供可靠的因果信息。



VLM 原始输出：[Robot arm][Lifting][Small black object][Upwards][One are's length]

译文：机械手臂举出黑色小物体向上一只手臂的长度



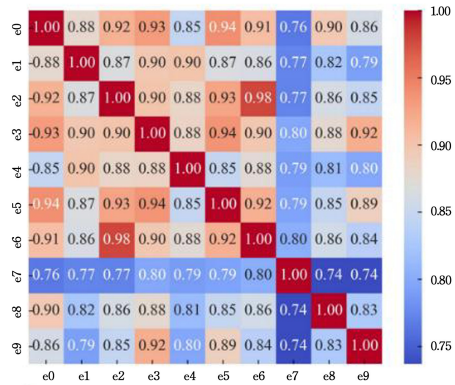
VLM 原始输出：[Robot arm][Grasping][Green ball][Towards the right][Small displacement, less than one arm's length]

译文：机械手臂抓握绿色球向右小位移，小于一只手臂的长度

图 3 多模态大模型生成语义动作的示例

Fig. 3 Examples of some MLM-generated actions

进一步地，为评估上下文量化编码的效果，对预训练数据集中的视频帧的上下文量化编码进行相关性分析。图 4 显示，相近类别视频的上下文量化编码具有较高的相似性。例如，在 SSv2 数据集中，“放下”(e0)和“举起”(e5)的编码的关联程度高达 0.95；而“浇”(e1)和“展示”(e9)等不相似场景的关联程度较低(0.79)。上下文量化编码有效体现了不同类别视频的相似度和区分度，避免了背景冗余干扰，增强了世界模型对复杂视频的理解。



e0：Moving something down

e1：Pouring something onto something

e2：Lifting something up completely, then letting it drop down

e3：Putting something infront of something

e4：Trying to pour something into something, but missing so it spills next to it

e5：Picking something up

e6：Lifting up one end of something, then letting it drop down

e7：Plugging something into something but pulling it right out as you remove your hand

e8：Spinning something soit continues spinning

e9：Showing something to the camera

图 4 采样部分 CVQ 编码之间的关系

Fig. 4 Relationship of sampled CVQ embeddings

结束语 本文提出了一种基于多模态大模型辅助视频动作生成的预训练世界模型 MAPO，预训练阶段无需显式动作标签，仅依赖互联网视频与多模态大模型生成伪动作，解除了传统预训练对标注数据的依赖；微调阶段通过融合多模态语义推理与动力学建模，显著提升了世界模型在视觉强化学习任务中的样本效率与泛化能力，进一步提升了下游任务中想象轨迹的准确性，以优化策略学习。性能评估证明，相比现有预训练世界模型方法，本文的 MAPO 方案在机器人运动与操作任务中均展现了显著优势。实验分析进一步验证了多模态生成动作的因果合理性与上下文编码的语义可分性，为未来探索开放场景下的通用世界模型提供了新的技术路径。

参 考 文 献

[1] MOERLAND T M, BROEKENS J, PLAAT A, et al. Model-based reinforcement learning: A survey[J]. Foundations and Trends © in Machine Learning, 2023, 16(1): 1-118.

[2] LUO F, XU T, LAI H, et al. A survey on model-based reinforcement learning[J]. Science China (Information Sciences), 2024(2): 067.

[3] HA D, SCHMIDHUBER J. Recurrent world models facilitate policy evolution[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 2455-2467.

[4] HAFNER D, LILLICRAP T P, BA J, et al. Dream to control: Learning behaviors by latent imagination[C]// International Conference on Learning Representations. 2020.

[5] HAFNER D, LILLICRAP T P, NOROUZI M, et al. Mastering atari with discrete world models[C]// International Conference on Learning Representations. 2021.

[6] HU A, RUSSELL L, YEO H, et al. GAIA-1: A generative world model for autonomous driving[J]. arXiv: 2309.17080, 2023.

[7] BAI C, XU H, LI X. Embodied-AI with large models: research and challenges[J]. Science China (Information Sciences), 2024, 54: 2035-2082.

[8] HANSEN N, SU H, WANG X. TD-MPC2: Scalable, robust world models for continuous control[C]// The Twelfth International Conference on Learning Representations. Vienna, Austria, 2024.

[9] XU Y, PARKER-HOLDER J, PACCHIANO A, et al. Learning general world models in a handful of reward-free deployments[J]. Advances in Neural Information Processing Systems, 2022, 35: 26820-26838.

[10] FENG Y, HANSEN N, XIONG Z, et al. Finetuning offline world models in the real world[C]// Conference on Robot Learning. PMLR, 2023: 425-445.

[11] SHAH S, DEY D, LOVETT C, et al. Airsim: High-fidelity visual and physical simulation for autonomous vehicles[C]// Field and Service Robotics: Results of the 11th International Conference. Springer International Publishing, 2018: 621-635.

[12] CHEN X, JIANG S, XU F, et al. Cross-modal domain adaptation for cost-efficient visual reinforcement learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 12520-12532.

[13] LIN Q，YU C，WU X，et al. Survey on Sim-to-real Transfer Reinforcement Learning in Robot Systems[J]. Journal of Software，2024，35(2)：711-738.

[14] MA W，LI S，CAI L，et al. Learning modality knowledge alignment for cross-modality transfer[C]// Proceedings of the 41st International Conference on Machine Learning. 2024：33777-33793.

[15] SEO Y，LEE K，JAMESS S L，et al. Reinforcement learning with action-free pre-training from videos[C]// Proceedings of International Conference on Machine Learning. PMLR，2022：19561-19579.

[16] ZHANG L，KAN M，SHAN S，et al. PreLAR：World model pre-training with learnable action representation[C]// European Conference on Computer Vision. Cham：Springer Nature Switzerland，2024：185-201.

[17] KINGMA D P，WELLING M. Auto-encoding variational bayes[C]// International Conference on Learning Representations. 2014.

[18] WU J，YIN S，FENG N，et al. iVideoGPT：Interactive videogpts are scalable world models[J]. Advances in Neural Information Processing Systems，2024，37：68082-68119.

[19] MICHELI V，ALONSO E，FLEURET F. Transformers are sample-efficient world models[C]// The Eleventh International Conference on Learning Representations. 2023.

[20] ZHANG W，WANG G，SUN J，et al. Storm：Efficient stochastic transformer based world models for reinforcement learning[J]. Advances in Neural Information Processing Systems，2023，36：27147-27166.

[21] ROBINE J，HOFTMANN M，UELWER T，et al. Transformer-based World Models Are Happy With 100k Interactions[C]// The Eleventh International Conference on Learning Representations. 2023.

[22] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Proceedings of the 31st Internatioan Conference on Neural Information Processing Systems. 2017：6000-6010.

[23] BELLEMARE M G，NADDAF Y，VENESS J，et al. The arcade learning environment：An evaluation platform for general agents[J]. Journal of Artificial Intelligence Research，2013，47：253-279.

[24] DENG F，PARK J，AHN S. Facing off world model backbones：Rnns，transformers，and s4[J]. Advances in Neural Information Processing Systems，2023，36：72904-72930.

[25] SOHL-DICKSTEIN J，WEISS E，MAHESWARANATHAN N，et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]// Proceedings of International Conference on Machine Learning. PMLR，2015：2256-2265.

[26] ALONSO E，JELLEY A，MICHELI V，et al. Diffusion for world modeling：Visual details matter in atari[J]. Advances in Neural Information Processing Systems，2024，37：58757-58791.

[27] DING Z，ZHANG A，TIAN Y，et al. Diffusion world model：Future modeling beyond step-by-step rollout for offline reinforcement learning[J]. arXiv：2402.03570，2024.

[28] WU J，MA H，DENG C，et al. Pre-training contextualized world models with in-the-wild videos for reinforcement learning[J]. Advances in Neural Information Processing Systems，2023，36：39719-39743.

[29] LU C，SCHROECKER Y，GU A，et al. Structured state space models for in-context reinforcement learning[J]. Advances in Neural Information Processing Systems，2023，36：47016-47031.

[30] KAELBLING L P，LITTMAN M L，CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence，1998，101(1/2)：99-134.

[31] QWEN TEAM. Qwen2.5-VL[EB/OL]. [2025-01-31]. https://qwenlm.github.io/blog/qwen2.5-vl/.

[32] VAN DEN OORD A，VINYALS O. Neural discrete representation learning[C]// NIPS. 2017.

[33] TASSA Y，DORON Y，MULDAL A，et al. Deepmind control suite[J]. arXiv：1801.00690，2018.

[34] YU T，QUILLEN D，HE Z，et al. Meta-world：A benchmark and evaluation for multi-task and meta reinforcement learning[C]// Conference on Robot Learning. PMLR，2020：1094-1100.

[35] GOYAL R，EBRAHIMI KAHOU S，MICHALSKI V，et al. The "something something" video database for learning and evaluating visual common sense[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：5842-5850.

[36] IONESCU C，PAPAVA D，OLARU V，et al. Human3.6m：Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，36(7)：1325-1339.

WAN Shenghua，born in 1999，doctoral candidate，is a student member of CCF (No. I7496G). His main research interests include reinforcement learning and world models.

ZHAN Dechuan，born in 1982，Ph. D，professor，is a member of CCF (No. 20015M). His main research interests include machine learning and data mining.

（责任编辑：柯颖）