DOI: 10.19650/j.cnki.cjsi.J2614971

基于三维几何世界模型的具身智能方法研究*

覃宏伟[1],姜 杨[2],白佳硕[2],徐天尧[2],杨轩溥[2]
(1. 东北大学信息科学与工程学院 沈阳 110169; 2. 东北大学机器人科学与工程学院 沈阳 110169)

摘 要:现有的具身智能操控方法大多采用端到端的学习策略,虽在特定任务上表现尚可,但因将物理交互视为黑箱问题,牺牲了系统可解释性,且面临几何感知缺失、长程规划累积误差大及动力学适应性差等挑战。为解决非结构化场景下的精细操作难题,提出了一种基于三维几何世界模型(3D-GWM)与条件流匹配策略的具身智能操控框架,包含三维感知与压缩、生成式几何世界模型和条件流匹配动作生成 3 个核心模块。感知模块提出了互补双域特征增强算法,综合提取空间高频细节与全局频域上下文,并结合指令驱动的动态词元压缩机制,在保证高保真特征提取的同时降低计算开销。部署评估中,输入视觉词元数由 2 048 压缩至 128,峰值显存占用由 58.7 GB 降至 17.2 GB。认知模块采用中间帧预测策略,通过预测子任务阶段性的三维几何终态,为长程任务提供具备因果逻辑的中间目标引导,有效抑制规划过程中的累积误差。执行模块采用条件流匹配策略,利用几何锚点构建目标向量场,生成符合物理接触约束的平滑动力学轨迹,降低刚性逆运动学求解中奇异性问题的发生风险,并缓解轨迹不连续带来的执行不稳定。实验结果表明,3D-GWM 在复杂逻辑与高精度任务中的仿真平均成功率为 89.4%,真机平均成功率为 82.0%;相较于当前代表性基线方法,平均成功率分别提升 7.0% 和 8.75%。

关键词: 具身智能;机器人操控;世界模型;条件流匹配;互补双域特征增强
中图分类号: TH86 TP242.6 **文献标识码**: A **国家标准学科分类代码**: 520.20 510.80

Research on embodied AI methods based on 3D geometric world models

Qin Hongwei[1], Jiang Yang[2], Bai Jiashuo[2], Xu Tianyao[2], Yang Xuanpu[2]
(*1. School of Information Science and Engineering, Northeastern University, Shenyang 110169, China;*
2. School of Robot Science and Engineering, Northeastern University, Shenyang 110169, China)

Abstract: Most existing embodied manipulation methods predominantly rely on end-to-end learning strategies. Although performing adequately on specific tasks, these approaches often treat physical interaction as a black box, thereby sacrificing interpretability and suffering from insufficient geometric grounding, large error accumulation in long-horizon planning, and weak adaptability to complex dynamics. To address fine-grained manipulation in unstructured scenarios, this article proposes an embodied manipulation framework based on a 3D geometric world model (3D-GWM) and conditional flow matching, termed 3D-GWM, which comprises three modules: 3D perception and compression, a generative geometric world model, and conditional flow-matching-based action generation. In the perception module, a complementary dual-domain feature enhancement algorithm is proposed to jointly capture spatial high-frequency details and global frequency-domain context, together with an instruction-driven dynamic token compression mechanism to reduce computation while preserving feature fidelity. In deployment evaluations, the number of input visual tokens is reduced from 2 048 to 128, while the peak GPU memory usage is reduced from 58.7 GB to 17.2 GB. In the cognition module, an intermediate-frame prediction strategy is adopted. By predicting stage-wise 3D geometric terminal states of sub-tasks, it provides intermediate goal guidance with causal logic for long-horizon manipulation and mitigates error accumulation during planning. In the execution module, a conditional flow matching strategy is employed. By using geometric anchors to construct target vector fields, it generates smooth trajectories that satisfy physical contact constraints and reduces the risk of singularities commonly encountered in rigid inverse-kinematics solvers. Experimental results show that 3D-GWM achieves an average success rate of 89.4% in simulation and 82.0% on the real robot for complex-logic and high-precision tasks; compared with representative baseline methods, the average success rate

收稿日期:2026-02-28 Received Date: 2026-02-28
*基金项目:工业机器人用具身智能大模型平台研究与示范应用(2024QK2001)项目资助

increases by 7.0% and 8.75%, respectively.

Keywords: embodied AI; robotic manipulation; world model; conditional flow matching; complementary dual-domain feature enhancement

0 引 言

具身智能作为人工智能通往物理世界的关键桥梁,其核心目标在于构建能够理解非结构化环境并执行精细操作的智能系统[1]。在此背景下,世界模型作为智能体的认知中枢占据着至关重要的地位,其负责模拟物理世界的演变规律并预测未来状态,从而为决策提供因果依据。随着具身视觉感知与三维场景表征技术的持续发展,赋予机器人对物理世界几何结构、空间关系与对象状态的感知与表征能力,已成为提升复杂操作任务性能的重要趋势[2-3]。与此同时,围绕机器人抓取姿态检测与三维目标表征的研究,为复杂场景下精细操作的几何感知与抓取决策提供了重要基础[4]。构建高效且精确的生成式几何世界模型,不仅能够帮助机器人理解复杂的空间拓扑关系,还能通过在潜在空间中推演操作结果来辅助长程任务规划,对于提升机器人在复杂动态环境下的操控水平具有重要意义。

随着大语言模型(large language model, LLM)与 ViT(vision Transformer)技术的深度耦合,基于学习的具身智能范式在近年来取得了显著进展[5]。以 RT-2[6]、OpenVLA[7]为代表的新一代视觉语言动作模型(vision-language-action model, VLA),通过在大规模互联网数据上的预训练,赋予了机器人较强的语义理解能力与泛化执行指令的能力。为了弥补二维视觉在空间感知上的不足,GeoVLA[8]通过引入点云编码器显式地增强了 VLA 模型对物理世界几何信息的捕获能力;而 PointVLA[9]等研究则进一步尝试将三维点云特征以免训练的方式注入视觉语言模型中,以实现更精准的空间推理与操控。这些研究共同表明,将高层语义推理与底层三维几何感知有机结合,显著提升了机器人对静态环境的理解能力。然而,缺失时间维度的瞬时感知难以洞悉物理世界的动态演化机理;因此,为了赋予智能体跨越时间维度的未来预演能力与规避潜在风险的因果直觉,引入生成式世界模型已成为提升复杂空间操控水平的重要趋势[10]。

当前主流的世界模型方法,如基于循环状态空间模型的 Dreamer[11]系列,以及以 Genie[12]为代表的基于视频生成的世界模型,大多沿用逐帧序列预测或高频状态更新的范式。这种高频的连续预测策略虽然在短期视频生成的连贯性上表现尚可,但在长程物理交互任务中,由于物理环境演变具有非线性及复杂的接触动力学特性,极易产生累积误差,使模型生成的未来状态偏离真实物理规律[13]。此外,这种微观层面的状态推演往往受限于局部视野,缺乏对任务阶段性宏观目标的把握,不仅造成了巨大的计算资源冗余,更难以直接指导高精度的机械臂动作规划[14]。

因此,如何在大模型语义理解的基础上,构建一种基于中间帧预测的世界模型,使得系统既能通过稀疏的阶段性状态生成提供长程因果引导,又能维持精确的三维空间约束以规避动力学误差,成为了当前具身智能操控领域亟待解决的关键问题。

基于此,为解决非结构化场景下的精细操作难题,本文提出了一种基于三维几何世界模型(3D geometric world model, 3D-GWM)与条件流匹配策略的具身智能操控框架。该框架的核心创新与主要贡献体现在 3 个核心模块:1)感知模块:提出了互补双域特征增强与指令驱动的动态词元压缩机制,在定向筛选关键视觉信息、保证高保真特征提取的同时显著提升了计算效率;2)认知模块:构建了生成式几何世界模型,通过推演阶段性三维几何终态的“中间帧预测”策略取代传统逐帧预测范式,有效阻断了物理环境演变及长程任务中的累积误差;3)执行模块:引入了基于条件流匹配的动作生成策略,利用预测的三维几何状态构建目标向量场,生成平滑且符合物理接触约束的动力学轨迹,降低了刚性逆运动学求解易产生的奇异性问题出现的风险。最后,在仿真实验平台与真机平台上开展的闭环操控实验表明,3D-GWM 表现出较强的几何推理与泛化能力,验证了该框架在复杂高精度任务中的有效性。

1 系统设计

1.1 系统整体架构

3D-GWM 是一个融合显式几何推理与生成式动作策略的具身智能操控框架,旨在通过构建对物理世界的因果认知来解决非结构化场景下的精细操作问题。如图 1 所示,系统整体架构主要由三维感知与压缩模块、生成式几何世界模型模块以及条件流匹配动作生成模块 3 个核心模块组成。

三维感知与压缩模块首先对输入的原始点云与任务指令进行语义分割及结构化体素化处理,初步构建稠密特征张量;随后利用互补双域特征增强算法进行特征增强;最终通过指令驱动的动态词元压缩机制,定向筛选并压缩出紧凑的三维语义词元序列。

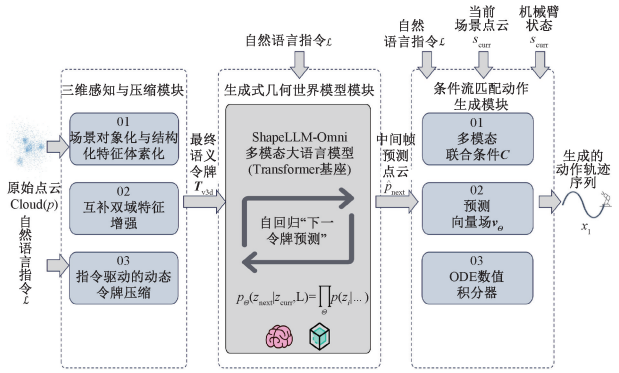

图 1　系统架构

Fig. 1　System architecture

生成式几何世界模型模块作为系统的认知中枢,接收来自感知模块的视觉词元与语言指令嵌入,并通过多模态大语言模型(multimodal large language model, MLLM)执行推理任务,即根据当前的物理状态,自回归地预测出下一个任务阶段的目标三维几何状态。

条件流匹配动作生成模块负责执行层面的动作求解,该模块将当前观测的点云状态与世界模型预测的目标几何状态作为联合条件输入,利用条件流匹配网络生成平滑、鲁棒且适应复杂接触交互的机械臂关节控制序列。

后文将具体介绍上述三大模块的实现方式,以及分段式训练策略的构建细节。

1.2　三维感知与压缩模块

三维感知与压缩模块负责将环境观测数据转换为兼具高频几何细节与全局语义的紧凑表征。如图 2 所示,该感知过程主要由 3 个阶段构成:阶段 1 为基于语义分割的场景对象化与结构化特征体素化,旨在构建保留三维空间拓扑连续性的稠密张量;阶段 2 为互补双域特征增强,通过三维卷积与频谱滤波在规则网格上并行提取局部纹理与全局形状特征;阶段 3 为指令驱动的动态词元压缩,依据自然语言意图从高维特征体中定向筛选关键语义信息。

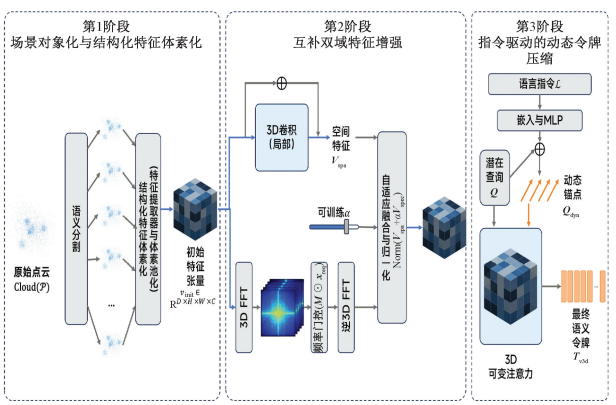

图 2　三维感知与压缩模型

Fig. 2　3D perception and compression model

1)场景对象化与结构化特征体素化

针对原始点云数据的无序性与稀疏性,本文引入预训练的语义分割模型(segment anything model, SAM)将全场景点云 $\mathcal{P}$ 解耦为 K 个语义独立的子集 $\mathcal{P}=\bigcup_{k=1}^{K}\mathcal{P}_k$。

鉴于直接坐标映射易导致的离散词元空间出现特征不连续问题,本文采用结构化特征体素化策略,将离散的点云数据 $\mathcal{P}_k$ 映射至规则的三维特征网格中,以构建稠密的特征张量,其具体实施过程为:首先,定义分辨率为 $D\times H\times W$ 的三维欧几里得网格空间,然后,利用点云特征提取器提取每个点的局部几何特征,并通过体素池化操作将点特征聚合至对应的网格单元中。

经此处理,非结构化的点云被重构为规则的四维特征张量 $\boldsymbol{V}_{\text{init}}\in\mathrm{R}^{D\times H\times W\times C}$,其中 C 为特征通道维度,从而确立了保留对象物理空间结构与相对尺度的张量数据基础。

2)互补双域特征增强

为解决三维体素在下采样过程中易丢失细粒度几何边缘且难以捕捉全局遮挡关系的问题,本文构建了基于规则网格的互补双域增强机制。该机制以特征张量 $\boldsymbol{V}_{\text{init}}$ 为输入,包含空间域与频域两条并行处理路径。

在空间域路径中,本文采用三维卷积对体素特征进行局部处理,以保留物体边缘与纹理等高频几何细节。引入残差连接结构后,空间增强特征 V_{spa} 计算如式(1)所示。

$$\boldsymbol{V}_{\text{spa}}=\text{Conv3D}(\boldsymbol{V}_{\text{init}})+\boldsymbol{V}_{\text{init}} \tag{1}$$

在频域路径中,本文先通过三维傅里叶变换将特征张量映射至频域空间,再利用可学习的频率门控矩阵 $\boldsymbol{M}$ 对频谱分量进行自适应调制,最后通过归一化融合空间与频域特征,得到增强特征 $\boldsymbol{V}_{\text{spec}}$。其计算过程如式(2)~(4)所示。

$$\boldsymbol{X}_{\text{freq}}=F_{3\text{D}}(\boldsymbol{V}_{\text{init}}) \tag{2}$$

$$\boldsymbol{V}_{\text{spec}}=F_{3\text{D}}^{-1}(\boldsymbol{M}\odot\boldsymbol{X}_{\text{freq}}) \tag{3}$$

$$\boldsymbol{V}_{\text{enh}}=\text{Norm}(\boldsymbol{V}_{\text{spa}}+a\boldsymbol{V}_{\text{spec}}) \tag{4}$$

其中,$\alpha\in[0,1]$ 为可训练参数,用于调节频谱域信息相对于空间域信息的贡献比例,从而在模型训练中平衡对高频细节的关注与对环境噪声的抑制。

3)指令驱动的动态词元压缩

为适配大语言模型有限的上下文窗口并提升推理效率,本文采用指令驱动的动态压缩算法将高维增强特征体 $\boldsymbol{V}_{\text{enh}}$ 转换为紧凑的词元序列。该过程利用自然语言指令 L 引导特征筛选。

首先,初始化一组与任务无关的潜在查询向量 $\boldsymbol{Q}\in\mathbf{R}^{M\times C}$,并通过多层感知机将语言指令嵌入注入查询中,生成蕴含特定操作意图的动态锚点 $\boldsymbol{Q}_{\text{dyn}}$,如式(5)所示。

$$\boldsymbol{Q}_{\mathrm{dyn}} = \boldsymbol{Q} + \mathrm{MLP}(\mathrm{Embed}(L)) \tag{5}$$

随后,本文将 $\boldsymbol{Q}_{\mathrm{dyn}}$ 作为查询向量,增强特征体 $\boldsymbol{V}_{\mathrm{enh}}$ 作为键值对,利用指令引导的交叉注意力机制在特征空间中搜索与任务相关的几何区域,实现了从稠密体素特征到稀疏语义词元的非线性映射,并输出最终的三维语义词元序列 $\boldsymbol{T}_{\mathrm{v3d}}$。 如式(6)所示。

$$\boldsymbol{T}_{\mathrm{v3d}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{dyn}}\boldsymbol{V}_{\mathrm{enh}}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_{\mathrm{enh}} \tag{6}$$

通过上述映射,冗余的体素数据被压缩为包含任务关键信息的高维向量序列,从而完成了从三维感知数据到认知推理输入的维度规约。

1.3 生成式几何世界模型模块

生成式几何世界模型模块作为系统的认知中枢,负责构建从自然语言指令到物理环境状态演变的因果映射。本文基于 ShapeLLM-Omni[15] 架构进行微调,利用其自回归生成特性与强大的 3D 编辑能力,将机器人操控任务重构为离散潜在空间内的多模态序列预测问题。具体推演流程如图 3 所示,主要包含多模态嵌入与统一离散表征、基于 MLLM 的中间状态推演与几何词元解码与度空间重构 3 个阶段。

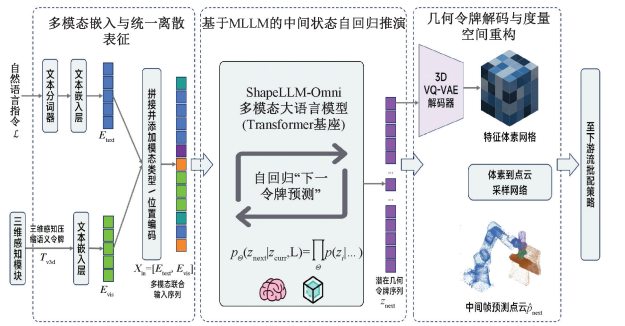

图 3 生成式几何世界模型模块

Fig. 3 Generative geometric world model module

1)多模态嵌入与统一离散表征

为打破文本与三维几何之间的模态壁垒,本文将异构数据统一映射至共享的离散词元空间。输入序列由自然语言指令 L 与三维感知模块输出的压缩语义词元 $\boldsymbol{T}_{\mathrm{v3d}}$ 构成。

首先,本文利用文本分词器将指令 L 转化为文本词元序列,并通过词嵌入层映射为文本嵌入 $\boldsymbol{E}_{\mathrm{text}}$。 针对视觉模态,鉴于 ShapeLLM-Omni 采用三维向量量化变分自编码器(3D vector quantized variational autoencoder, 3D VQ-VAE)代码本表示几何信息,本文设计了一个特征对齐适配器,将上游感知的压缩特征 $\boldsymbol{T}_{\mathrm{v3d}}$ 投影至 MLLM 的输入维度,使其语义分布与预训练模型的 3D 词元空间对齐,生成视觉嵌入 $\boldsymbol{E}_{\mathrm{vis}}$。 随后,本文引入可学习的模态类型编码与绝对位置编码,将 $\boldsymbol{E}_{\mathrm{text}}$ 与 $\boldsymbol{E}_{\mathrm{vis}}$ 拼接为多模态联合输入序列 $\boldsymbol{X}_{\mathrm{in}} = [\boldsymbol{E}_{\mathrm{text}}, \boldsymbol{E}_{\mathrm{vis}}]$,在输入层完成语义意图与当前几何表征的结构化统一。

2)基于多模态大模型的中间状态自回归推演

本文利用 MLLM 强大的通用推理能力作为几何引擎,通过微调使 MLLM 习得短时域内的物理状态转移规律,实现从当前观测到未来中间帧的因果推演,为动作生成提供精准的几何引导。

本文将中间帧定义为专家轨迹中相邻子任务之间的阶段关键帧,即能够表征阶段性三维几何终态并为后续动作生成提供中间目标引导的观测帧,其标注采用半自动构造流程。先对少量轨迹人工标注中间帧,再训练中间帧预测器,根据几何变化、接触事件及执行器状态实现其余轨迹的自动分段并生成中间帧,最后仅对困难样本进行人工抽查修正。

MLLM 接收多模态联合输入 $\boldsymbol{X}_{\mathrm{in}}$, 深度解析自然语言指令中蕴含的空间操作逻辑与当前几何状态的关联。其推理过程遵循“下一个词元预测”范式,模型根据历史上下文信息,自回归地预测表征下一阶段中间状态的离散几何词元序列,如式(7)所示。

$$p_\theta(\boldsymbol{z}_{\mathrm{next}} \mid \boldsymbol{z}_{\mathrm{curr}}, L) = \prod_{i=1}^{T} p_\theta(z_i \mid z_{<i}, \boldsymbol{z}_{\mathrm{curr}}, L) \tag{7}$$

在此过程中,模型执行显式的体积推理,利用预训练阶段在海量 3D 数据集中习得的物理常识与空间结构先验,在潜在空间中生成符合指令预期且具备物理合理性的中间态几何结构编码。

3)几何词元解码与度量空间重构

模型输出的潜在几何词元序列需映射回三维欧几里得空间,以作为下游条件流匹配策略的显式条件。为此本文采用 ShapeLLM-Omni 的 3D VQ-VAE 解码器架构,将预测的离散词元序列 $\boldsymbol{z}_{\mathrm{next}}$ 首先还原为结构化的特征体素网格。随后,利用体素到点云的采样网络,将体素网格解构并重构为中间帧预测点云 $\hat{S}_{\mathrm{next}}$。 该点云数据包含目标物体在下一阶段目标几何状态下的精确位姿与形态信息,其作为几何锚点被输入至后续的条件流匹配动作生成模块,从而实现从高层语义规划到底层几何约束的有效传递。

1.4 条件流匹配动作生成模块

本文在动作生成模块中构建了一个包含本体感知约束的多模态条件流匹配[16]框架,如图 4 所示,主要包含多模态联合条件流匹配网络的构建以及流式轨迹生成的微分方程求解两个阶段,旨在解决高维连续动作空间内的逆运动学求解问题。

该模块以世界模型生成的未来几何状态为引导,通过学习从当前机械臂状态到下一阶段预测状态的概率流向量场,在连续时间域上通过常微分方程积分求解出符合运动学约束的动作序列。

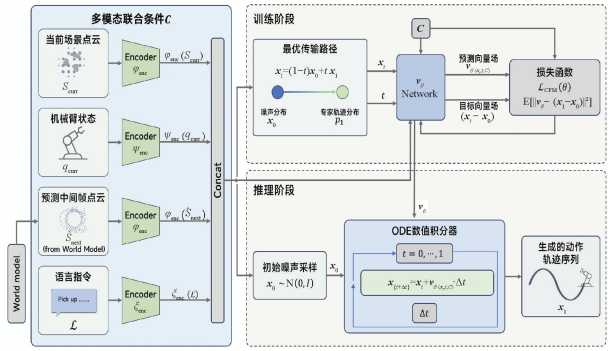

图 4　条件流匹配动作生成模块

Fig. 4　Conditional flow matching action generation module

1)多模态联合条件流匹配网络构建

为确保生成的动作轨迹兼顾环境几何约束与机械臂自身的运动学状态,本文构建了一个深度神经网络作为时变向量场估计器 $\boldsymbol{v}_\theta$。与传统仅依赖视觉输入的策略不同,该网络的输入条件向量 C 被设计为多模态特征的联合表征,如式(8)所示。

$$\boldsymbol{C} = \mathrm{Concat}(\phi_{\mathrm{enc}}(S_{\mathrm{curr}}),\psi_{\mathrm{enc}}(\boldsymbol{q}_{\mathrm{curr}}),\varphi_{\mathrm{enc}}(\hat{S}_{\mathrm{next}}),\xi_{\mathrm{enc}}(L)) \tag{8}$$

其中,$\phi_{\mathrm{enc}}(S_{\mathrm{curr}})$ 为当前时刻场景点云的几何特征,用于表征环境障碍物分布;$\psi_{\mathrm{enc}}(q_{\mathrm{curr}})$ 为机械臂当前的关节角度与角速度编码,作为动作生成的物理锚点;$\varphi_{\mathrm{enc}}(\hat{S}_{\mathrm{next}})$ 为世界模型直接预测的中间帧点云特征,提供显式的几何目标引导;$\xi_{\mathrm{enc}}(L)$ 为语言指令嵌入,用于在高层语义上消除动作歧义。

在训练阶段,该网络旨在回归一个能够将初始噪声分布 p_0 映射到真实专家轨迹分布 p_1 的向量场。为提高生成效率与训练稳定性,本文采用最优传输路径作为监督信号,构建直线概率流。对于任意时间步 $t \in [0,1]$,中间状态 $\boldsymbol{x}_t$ 由专家轨迹 $\boldsymbol{x}_1$ 与噪声 $\boldsymbol{x}_0$ 线性插值得到,即 $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\cdot\boldsymbol{x}_1$。网络的优化目标定义为最小化预测向量场与目标条件向量场之间的期望均方误差,如式(9)所示。

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathrm{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1}[\,\|\boldsymbol{v}_\theta(\boldsymbol{x}_t,t,\boldsymbol{C}) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2] \tag{9}$$

通过该损失函数,网络学习到了在多模态条件 $\boldsymbol{C}$ 的直接引导下,驱动系统状态从无序噪声向确定性专家动作演变的动力学规律,使其能够适应世界模型预测的几何特征分布。

2)流式轨迹生成的微分方程求解

在推理阶段,动作轨迹的生成过程被建模为常微分方程的初值求解问题。本文直接利用训练好的向量场网络 v_θ,在当前观测与世界模型预测的中间帧点云 $\hat{S}_{\mathrm{next}}$ 的联合控制下,计算流随时间演变的瞬时速度。

首先,本文从标准高斯分布中采样随机噪声向量 $\boldsymbol{x}_0$ 作为初始流状态。随后,本文利用数值积分器在时间域上进行迭代推进。对于步长 Δt,状态更新遵循以下动力学方程,如式(10)所示。

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \boldsymbol{v}_\theta(\boldsymbol{x}_t,t,\boldsymbol{C})\cdot\Delta t \tag{10}$$

随着时间 t 从 0 推进至 1,初始噪声逐步演化为确定性的动作轨迹序列 $\boldsymbol{x}_1$。在此过程中,该条件流匹配网络发挥了其强大的分布映射能力,能够鲁棒地处理预测点云中的几何信息,将其映射为符合机械臂运动学约束的平滑轨迹,从而实现了从几何意图到物理执行的端到端生成。

1.5　训练策略

3D-GWM 系统涉及视觉感知、几何认知与动作执行等异构模态。为规避端到端联合训练中易产生的梯度冲突与模态坍塌问题,本文采用“感知–认知–执行”的三阶段渐进式训练范式。该策略在对齐各模态特征空间的同时,有效隔离不同任务的优化目标。其具体流程如图 5 所示。

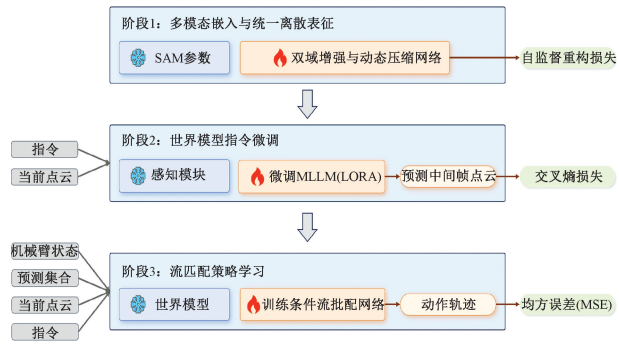

图 5　训练策略示意图

Fig. 5　Schematic of training strategy

1)阶段 1:三维感知适配层预热

本阶段旨在启动三维感知与压缩模块的特征提取能力。在冻结底层语义分割模型参数的基础上,本文仅更新互补双域特征增强网络及动态词元压缩器的权重。同时,本文构建自监督重构任务,通过最小化特征重构损失,驱动稀疏语义词元 $\boldsymbol{T}_{\mathrm{v3d}}$ 在压缩状态下最大化保留原始场景的高频几何细节与语义拓扑,从而为世界模型提供高信噪比的结构化输入。

2)阶段 2:世界模型指令微调

本阶段旨在赋予认知中枢从语言到几何的短时域推演能力。在冻结感知模块参数的前提下,本文采用低秩适配技术(low-rank adaptation, LoRA)微调 MLLM 注意力层。为构建“指令—当前点云—中间帧点云”三元组,本文首先在少量专家轨迹上人工标注中间帧,并据此训练中间帧预测器;随后利用该预测器对其余轨迹进行自动分段,生成阶段性中间帧监督信号,并仅对少量困难样本

进行人工抽查修正。在此基础上,本文将优化目标设定为最小化中间帧几何词元序列的预测交叉熵损失,从而使模型习得根据当前观测与任务指令预测下一阶段几何状态的能力。

3)阶段 3:条件流匹配策略学习

本阶段基于世界模型的几何预测训练执行层的条件流匹配动作生成模型。其输入条件为当前场景特征、机械臂关节状态、语言指令与预测中间帧特征的联合向量。在此阶段,本文利用专家轨迹数据构建向量场回归任务,通过最小化预测向量场与目标向量场之间的均方误差,使条件流匹配网络在多模态特征引导下,学习从标准高斯噪声向平滑、可执行的关节控制轨迹演变的概率流规律。

2 实验与结果分析

2.1 实验环境与设置

为全面评估所提出的 3D-GWM 架构在具身智能操控任务中的性能,本文开展了多场景、多任务设置下的仿真与真机实验,旨在从几何预测的准确性、动作生成的鲁棒性等维度,验证“感知—认知—执行”分层架构的有效性。

1)仿真实验平台

本研究选用 LIBERO[17] 作为仿真实验平台,如图 6 所示。LIBERO 是基于 multi-joint dynamics with contact (MuJoCo)物理引擎构建的仿真实验平台,广泛应用于评估具身智能策略的空间推理与泛化能力。

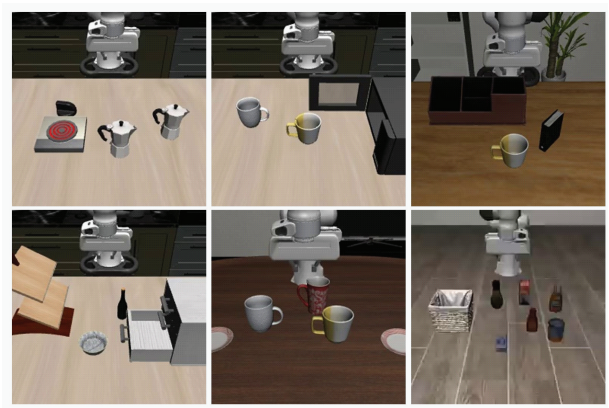

图 6 仿真实验平台

Fig. 6 Experimental setup in the LIBERO simulation environment

为了全方位评估模型性能,本文具体选取了 LIBERO-Spatial、LIBERO-Object、LIBERO-Goal 以及 LIBERO-Long 这 4 个子集。其中,LIBERO-Spatial 侧重于检验模型对空间方位与几何关系的理解能力;LIBERO-Object 侧重于评估模型对不同物体形状与纹理的适应性;LIBERO-Goal 侧重于测试模型对特定目标指令的理解与执行准确度;而 LIBERO-Long 则包含长程多阶段任务,旨在重点验证本文提出的中间帧预测策略在抑制长序列规划累积误差方面的有效性。

2)真机实验平台

本文的硬件执行系统如图 7 所示,采用 RealMan RM-65 六自由度机械臂,末端搭载平行夹爪;视觉感知系统采用单台 RealSense D435 深度相机,置于平台顶部俯瞰全局;所有训练与推理任务均部署于配备 NVIDIA H100 GPU 的服务器中。

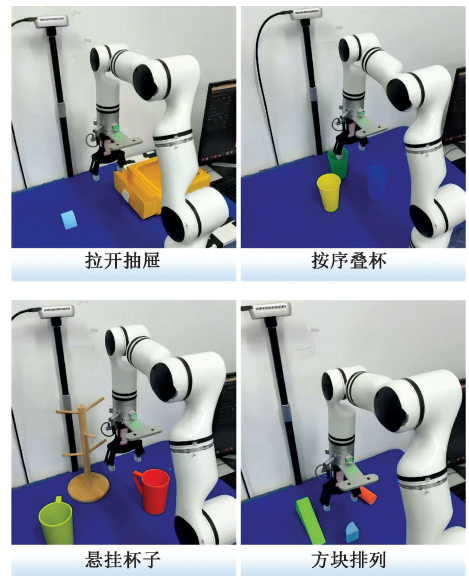

图 7 真机实验平台

Fig. 7 Real-world robotic experimental platform

为了验证 3D-GWM 架构在真实物理环境中的综合性能,本文选取如图 7 所示的 4 项典型任务分别进行 100 次测试。其中,“拉开抽屉”涉及复杂的物理接触,重点评估模型在丰富场景下的动力学适应性;“按序叠杯”需严格依照颜色顺序堆叠,重点验证多阶段长程任务规划能力;“悬挂杯子”要求将杯柄挂入挂钩,重点测试模型动作精度;“方块排列”则通过对不同属性积木的有序排列,综合探究模型对物体属性的语义理解与空间逻辑推理能力。

3)算法实现细节

3D-GWM 基于 PyTorch 深度学习框架构建,运行环境为 Ubuntu 22.04 操作系统配合 CUDA 12.1 加速库,其训练过程严格遵循前文所述的三阶段分段式策略。在第 1 阶段感知适配预热中,本文使用 AdamW[18] 优化器训练互补双域特征增强网络,初始学习率设为 2×10^{-4},并采用余弦退火策略[19] 进行衰减,Batch Size 设为 32。在第 2 阶段世界模型微调中,基于预训练的 ShapeLLM-

Omni 骨干网络,本文通过 LoRA[20] 技术对注意力层的投影矩阵进行低秩更新,其中 LoRA rank 设为 64,alpha 设为 16,训练窗口长度为 512 个词元。在第 3 阶段条件流匹配策略学习中,条件流匹配网络采用 diffusion transformer(DiT)架构,时间步 $t \sim U(0,1)$,推理时采用 Euler 数值求解器进行 10 步积分以生成动作轨迹。为保证实验的可复现性,本文将所有随机种子均固定为 42。在第 2 阶段数据构造中,中间帧监督信号采用少量人工标注结合中间帧预测器自动扩展的半自动方式获得,并对自动分段结果进行抽样人工复核。

2.2　对比试验及结果分析

为了验证 3D-GWM 方法的有效性和泛化性,本文以任务成功率作为核心评价指标,选取 4 种当前主流具身智能策略作为基线方法进行对比,包括 WorldVLA[21]、OpenVLA[7]、Diffusion Policy[22] 以及 DiT Policy[23]。为确保对比公平性,本文将所有基线模型均在相同的数据集与硬件环境下进行重新训练或微调。

1)仿真实验平台测试结果及分析

不同方法在 LIBERO 仿真实验平台上的成功率结果如表 1 所示,核心仿真实验结果如图 8 所示。

表 1　不同方法在 LIBERO 仿真实验平台上的成功率对比

Table 1　Comparison of success rates on LIBERO benchmark　(%)

模型	LIBERO-Spatial	LIBERO-Object	LIBERO-Goal	LIBERO-Long	平均成功率
WorldVLA[21]	87.60	96.20	83.40	60.00	81.80
OpenVLA[7]	84.70	88.40	79.20	53.70	76.50
Diffusion Policy[22]	78.30	92.50	68.30	50.50	72.40
DiT Policy[23]	84.20	96.30	85.40	63.80	82.40
3D-GWM (本文方法)	**92.40**	**96.60**	**85.20**	**83.40**	**89.40**

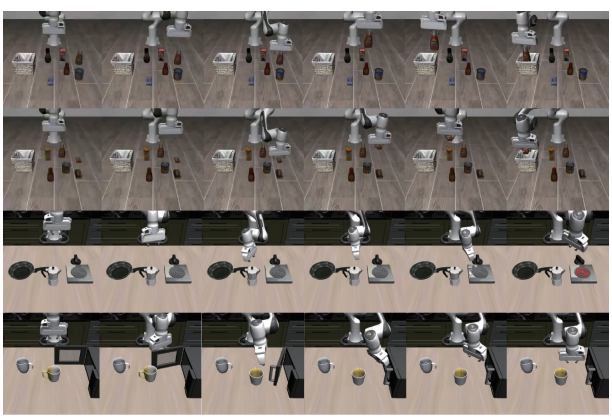

图 8　仿真平台典型任务演示

Fig. 8　Task demonstrations on the LIBERO simulation platform

从整体结果来看,3D-GWM 以 89.40% 的平均成功率取得总体最佳性能,明显高于 DiT Policy 与 WorldVLA,表现出稳定的综合优势。在 LIBERO-Spatial 任务中,本文方法达到 92.40% 的成功率。该类任务对空间结构理解与几何关系建模要求较高,结果表明所提框架能够更加准确地捕获三维几何拓扑特征,从而提升空间相关操作的执行稳定性。在最具挑战性的 LIBERO-Long 长程任务中,本文方法取得 83.40% 的成功率,显著高于其他对比方法。这一性能提升主要归因于生成式几何世界模型所引入的中间帧预测机制,其提供的阶段性几何引导能够有效抑制长序列规划过程中的累积误差。此外,在 LIBERO-Object 任务中,本文方法同样取得最高结果;在 LIBERO-Goal 任务中,3D-GWM 取得 85.20% 的成功率,略低于 DiT Policy 的 85.40%,但仍保持与最优基线接近的竞争性能。综合来看,所提方法在多类任务场景下均展现出较强的泛化能力与稳定性。

2)真机操控性能评估

真机环境下的实验结果如表 2 所示,核心真机实验演示如图 9 所示,3D-GWM 在真实且复杂的物理交互环境中平均成功率达到 82.00%,高于次优方法 DiT Policy。值得注意的是,在对末端位姿精度要求极高的“悬挂杯子”任务中,本文方法取得了 81.00% 的成功率,显著高于 WorldVLA 与 OpenVLA。这一显著差距主要归因于上述基线模型受限于离散化动作策略,在精细操作中不可避免地产生量化误差;而 3D-GWM 基于条件流匹配策略生成连续且平滑的动力学轨迹,有效规避了动作僵硬与奇异性问题。此外,在“拉开抽屉”与“按序叠杯”任务中,本文方法分别达到 73.00% 与 92.00% 的高成功率,充分验证了基于中间帧预测的世界模型在处理复杂场景与长程时序逻辑任务时的鲁棒性。

表 2　真机环境下典型操控任务的成功率对比

Table 2　Success rate comparison of typical manipulation tasks in real-world environment　(%)

方法	拉开抽屉	按序叠杯	悬挂杯子	方块排列	平均成功率
WorldVLA[21]	55	88	42	76	65.25
OpenVLA[7]	50	84	36	74	61.00
Diffusion Policy[22]	48	80	60	50	59.50
DiT Policy[23]	62	88	71	72	73.25
3D-GWM (本文方法)	**73**	**92**	**81**	**82**	**82.00**

2.3　消融实验及结果分析

为系统性验证 3D-GWM 框架中各组件的有效性,本文分别对感知、认知与执行层进行了消融研究。实验结

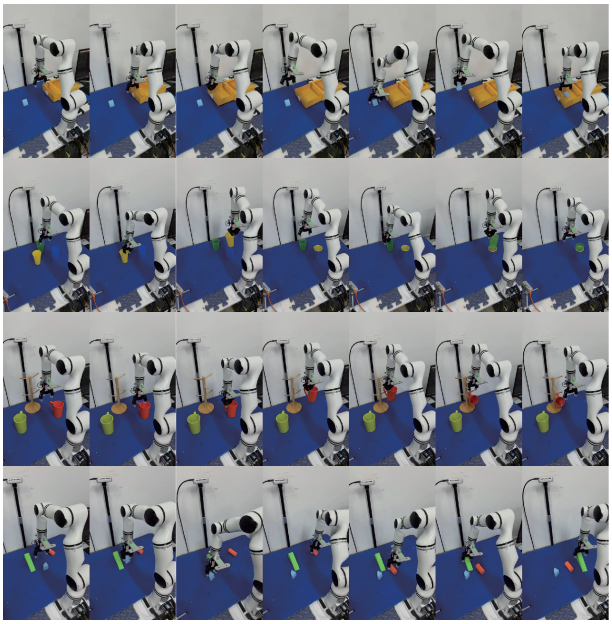

图 9 真机平台典型任务演示

Fig. 9 Representative real-world manipulation tasks on the robotic platform

果表明,感知层的互补双域特征增强与动态词元压缩,以及认知层的几何因果推理显著提升了模型在复杂交互任务中的鲁棒性,而执行层的条件流匹配策略则有效保障了长程动作生成的连贯性与精度。各模块的协同作用显著提升了任务成功率。

1)感知层消融实验及分析

如表 3 的定量结果表明,3D-GWM 在所有任务中均表现出优于 Baseline 的性能,充分验证了互补双域特征增强与指令驱动的动态词元压缩机制的有效性。去除互补双域特征增强(without dual-domain enhancement, w/o DD)后,模型在几何敏感任务中的表现明显下降,“拉开抽屉”与“悬挂杯子”的任务成功率分别由 73% 下降至 61% 和 62%,说明该模块有助于捕捉细粒度空间拓扑结构,从而提升高精度操作的稳定性。去除指令驱动的动态词元压缩机制(without instruction-driven dynamic token compression, w/o DTC)后,“按序叠杯”与“方块排列”任务成功率分别由 92% 下降至 80% 和 68%,表明该机制能够依据任务语义筛选关键视觉信息,从而增强模型对目标属性与空间关系的判别能力。综合来看,两种机制具有明显的互补作用,其联合使用时取得了最高的平均成功率 82.0%。

表 3 感知层消融实验结果

Table 3 Ablation study results of the perception layer

(%)

实验设置	双域增强	动态压缩	拉开抽屉	按序叠杯	悬挂杯子	方块排列	平均成功率
Baseline	×	×	58	80	52	61	62.75
w/o DTC	√	×	68	80	66	68	70.50
w/o DD	×	√	61	88	62	75	71.50
本文方法	√	√	**73**	**92**	**81**	**82**	**82.00**

2)认知层消融实验及分析

如表 4 所示,3D-GWM 在各项任务中均明显高于端到端模型(E2E-Flow)与逐帧预测模型(Seq-WM)。

表 4 认知层与执行层消融实验结果

Table 4 Ablation study results of cognition and execution layers

(%)

模块	模型变体	认知层	执行层	拉开抽屉	按序叠杯	悬挂杯子	方块排列	平均成功率
认知层	Seq-WM	逐帧预测	逆运动学	65	72	70	76	70.75
认知层	E2E-Flow	端到端预测	条件流匹配策略	63	85	72	76	74.00
执行层	Key-Diff	中间帧预测	扩散策略	68	88	74	79	77.25
执行层	Key-IK	中间帧预测	逆运动学	12	8	5	15	10.00
Full Model	3D-GWM	中间帧预测	条件流匹配策略	**73**	**92**	**81**	**82**	**82.00**

具体而言,E2E-Flow 在接触丰富的“拉开抽屉”任务中成功率下降较为明显,说明仅依赖端到端映射难以稳定处理复杂物理交互过程中的几何约束与阶段切换。与此同时,Seq-WM 在长程“按序叠杯”任务中相较完整模型存在较大差距,这一现象主要归因于逐帧预测在长序列决策中容易累积误差,进而影响整体任务完成质量。相比之下,本文采用的中间帧预测策略通过提供稀疏且精确的阶段性几何终态,使模型能够在长时域任务中维持更好的规划连贯性与执行鲁棒性,从而显著提升复杂任务的整体成功率。

3)执行层消融实验及分析

执行层消融实验结果进一步表明,不同动作生成策

略对最终操控性能具有显著影响。具体来看,Key-IK 变体表现出明显的性能退化,其在所有任务中的成功率均处于较低水平。这主要是因为当前观测帧与预测中间帧之间往往存在较大的状态跨度,在不引入额外稠密路径规划约束的情况下,传统逆运动学方法难以直接求解出稳定、平滑且可执行的过渡轨迹,容易受到奇异位形、关节限位及局部不可达区域的影响。相比之下,Key-Diff 虽然在多数任务中取得了较好的结果,但在“悬挂杯子”等高精度任务中仍逊于完整模型。上述结果表明,本文采用的条件流匹配策略在轨迹平滑性、末端控制精度与跨阶段动作一致性方面具有更强优势,从而能够更有效地适配由中间帧引导的大跨度动作生成过程。

2.4　模型部署效率评估及结果分析

具身智能模型在真实物理环境中的闭环应用不仅依赖高精度的空间推理能力,亦受限于严格的计算开销与推理延迟。为量化所提 3D-GWM 架构的部署可行性,本文对比了 3D-GWM 完整架构与 w/o DTC 的工程部署效率,实验结果如表 5 及图 10 所示。

表 5　3D-GWM 模型变体计算开销对比

Table 5　Comparison of computational overhead for 3D-GWM model variants

模型变体 架构设定	输入视觉词 元数	峰值显存 占用/GB	真机平均 成功率/%
w/o DTC	2 048	58.7	70.50
3D-GWM	**128**	**17.2**	**82.00**

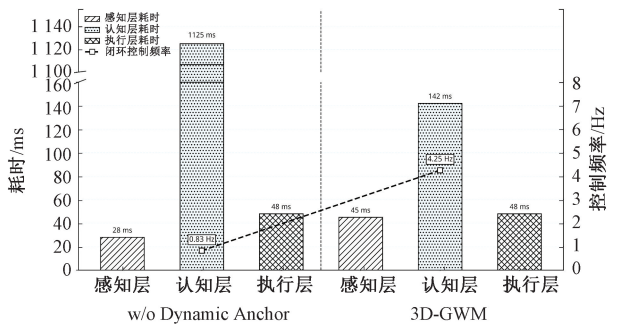

图 10　3D-GWM 模型变体推理延迟对比

Fig. 10　Inference latency comparison of 3D-GWM model variants

实验结果表明,该动态压缩机制在打破三维视觉信息瓶颈与降低算力开销中发挥了关键作用。当移除该机制时,三维非结构化场景经体素化产生 2 048 个稠密视觉词元,致使大语言模型注意力机制的计算负荷显著增加,峰值显存占用达 58.7 GB;其感知层耗时为 28 ms,但海量词元在认知层引发了耗时 1 125 ms 的自回归解码延迟,导致端到端闭环控制频率仅为 0.83 Hz,真机平均成功率为 70.50%,无法满足物理交互的实时性需求。相比之下,完整版 3D-GWM 以自然语言指令为查询向量进行靶向特征筛选,在保留高频几何细节的同时将高价值语义词元压缩至 128 个,有效缓解了键值缓存压力,将峰值显存降至 17.2 GB,为系统并发中间件预留了计算空间;同时,背景噪声的过滤减轻了多模态注意力的稀释效应,使真机平均成功率提升至 82.00%。

在模块级耗时分析中,完整版 3D-GWM 感知层耗时虽增至 45 ms,但认知层推理耗时大幅缩减至 142 ms。此外,执行层的条件流匹配网络在构建目标向量场及利用常微分方程求解动力学轨迹时表现出极高的稳定性,耗时恒定为 48 ms。

综合各模块性能,完整版 3D-GWM 最终实现了 4.25 Hz 的闭环控制频率,有效保障了机器人在动态物理交互过程中的响应平滑性,充分验证了该三维生成式世界模型实时真机部署的工程可行性。

2.5　真机失败案例分析

为更全面评估 3D-GWM 在真实物理环境中的系统边界,本文进一步对真机失败案例进行了统计与归因分析,如表 6 所示。

表 6　真机失败案例归因分析

Table 6　Attribution analysis of failure cases in real-world experiments

任务	总测试 次数	成功率/ %	失败 次数	主要 失败归因	典型 失败现象
拉开抽屉	100	73	27	感知受限、执行过渡误差	抓取偏移、拉动失稳
按序叠杯	100	92	8	中间帧偏差、执行过渡误差	叠放错位、放置不稳
悬挂杯子	100	81	19	中间帧偏差、执行过渡误差	对准不足、挂接失败
方块排列	100	82	18	感知受限、中间帧偏差	排列偏差、顺序混乱
合计	400	82	72	—	遮挡敏感、接触不稳

4 项任务共进行 400 次测试,其中成功 328 次、失败 72 次。结合任务执行现象可知,失败案例并非由单一模块引起,而主要表现为 3 类:1)感知受限,即遮挡、细粒度边缘或局部接触区域的关键几何信息保留不足,导致目标物体与操作区域的结构表征不完整;2)中间帧偏差,即生成式几何世界模型预测的阶段性三维几何终态与真实可执行状态之间存在偏离,尤其在高精度挂接与对准任

务中更为明显;3)执行过渡误差,即由当前状态向预测中间目标过渡时,末端姿态误差在接触瞬间被放大,进而导致抓取滑移、挂接失败或最终放置偏差。

总体而言,3D-GWM 的失败边界主要集中在遮挡条件下的精细感知、接触敏感任务中的中间目标预测精度以及大跨度动作过渡的稳定性上。

3 结 论

针对非结构化场景精细操控中几何约束不足、长程误差累积与接触轨迹生成不稳定等问题,本文提出了一种融合三维几何世界模型与条件流匹配的操控框架(3D-GWM)。在感知模块,本文提出了互补双域特征增强方法与指令驱动的动态词元压缩机制,以确保三维语义表征的紧凑性和保真性,为后续推演与控制提供稳定输入;在认知模块,本文构建生成式几何世界模型,并通过中间帧几何终态预测提供阶段性中间目标,以缓解长程规划误差累积;在执行模块,本文基于条件流匹配并结合几何锚点生成连续关节轨迹,从而提升接触过程的轨迹平滑性与可行性,并降低刚性逆运动学求解的奇异风险。本文在 LIBERO 仿真实验平台与 RealMan 真机平台上开展闭环实验,结果表明该框架在复杂逻辑任务与高精度操作任务上取得较好的性能,其中仿真实验平均成功率为 89.4%,真机实验平均成功率为 82.0%;相较代表性基线方法,仿真与真机平均成功率分别提升 7.0% 和 8.75%。此外,动态词元压缩机制将输入视觉词元数由 2 048 压缩至 128,并将峰值显存占用由 58.7 GB 降低至 17.2 GB,从而提升了模型的部署可行性。

但本文方法仍存在一定局限。首先,当前几何世界模型仍以分段短程预测为主,虽然能够在一定程度上缓解长程任务中的误差累积,但对跨阶段任务的全局时序记忆与长期依赖建模能力仍然不足。其次,本文实验主要集中于固定工作空间内的抓取、接触与装配类任务,对于开放环境下的动态干扰、场景变化以及移动操作等更复杂情形,其泛化能力仍有待进一步验证。此外,执行层动作生成依赖于世界模型提供的中间几何状态,当预测结果出现偏差时,误差可能进一步传递至后续动作规划与控制过程,从而影响执行稳定性与最终任务成功率。未来工作将围绕以下 3 个方向展开:1)引入显式长时序记忆机制或层次化状态摘要方法,以增强模型对长程任务全局结构的建模能力;2)将当前框架扩展到开放环境与移动操作场景,提升系统在复杂动态条件下的适应性;3)研究世界模型预测与执行策略之间的闭环校正机制,以降低中间状态误差对动作生成过程的累积影响,进一步提升系统鲁棒性与工程可部署性。

参考文献

[1] ZHENG Y, YAO L, SU Y J, et al. A survey of embodied learning for object-centric robotic manipulation[J]. Machine Intelligence Research, 2025, 22(4): 588-626.

[2] WANG S CH, NIKOLIĆ M N, LAM T L, et al. Robot manipulation based on embodied visual perception: A survey [J]. CAAI Transactions on Intelligence Technology, 2025, 10(4): 945-958.

[3] 葛俊彦, 史金龙, 周志强, 等. 基于三维检测网络的机器人抓取方法[J]. 仪器仪表学报, 2021, 42(8): 146-153.
GE J Y, SHI J L, ZHOU ZH Q, et al. A robotic grasping method based on three-dimensional detection network[J]. Chinese Journal of Scientific Instrument, 2021, 42(8): 146-153.

[4] 李秀智, 李家豪, 张祥银, 等. 基于深度学习的机器人最优抓取姿态检测方法[J]. 仪器仪表学报, 2020, 41(5): 108-117.
LI X ZH, LI J H, ZHANG X Y, et al. Detection method of robot optimal grasp posture based on deep learning[J]. Chinese Journal of Scientific Instrument, 2020, 41(5): 108-117.

[5] 邓鹏, 唐文涛, 罗静. 机器人大模型发展与挑战[J]. 电子测量与仪器学报, 2024, 38(12):12-25.
DENG P, TANG W T, LUO J. Robotic large model development and challenges[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(12): 12-25.

[6] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control[J]. ArXiv preprint arXiv: 2307.15818, 2023.

[7] KIM M J, PERTSCH K, KARAMCHETI S, et al. OpenVLA: An open-source vision-language-action model[J]. ArXiv preprint arXiv: 2406.09246, 2024.

[8] SUN L, XIE B, LIU Y F, et al. GeoVLA: Empowering 3D representations in vision-language-action models[J]. ArXiv preprint ArXiv:2508.09071, 2025.

[9] LI CH M, WEN J J, PENG Y X, et al. PointVLA: Injecting the 3D world into vision-language-action models[J]. IEEE Robotics and Automation Letters, 2026, 11(3): 2506-2513.

[10] DING J T, ZHANG Y K, SHANG Y, et al. Understanding world or predicting future? A comprehensive survey of world models[J]. ACM Computing Surveys, 2025, 58(3): 57:1-38.

[11] HAFNER D, PASUKONIS J, BA J, et al. Mastering

diverse control tasks through world models[J]. Nature, 2025, 640(8059): 647-653.
[12] BRUCE J, DENNIS M, EDWARDS A, et al. Genie: Generative interactive environments [C]. 2024 International Conference on Machine Learning, Part 6 of 75, 2024, 235: 4603-4623.
[13] TSUJI T, KATO Y, SOLAK G, et al. A survey on imitation learning for contact-rich tasks in robotics[J]. ArXiv preprint arXiv: 2506.13498,2025.
[14] 吴海波, 杨梦琦, 杨宇恒, 等. 基于数字孪生的机械臂路径规划研究[J]. 仪器仪表学报, 2026, 47(2): 173-185.
WU H B, YANG M Q, YANG Y H, et al. Research on path planning of robotic arm based on digital twin[J]. Chinese Journal of Scientific Instrument, 2026, 47(2): 173-185.
[15] YE J L, WANG ZH Y, ZHAO R W, et al. ShapeLLM-Omni: A native multimodal LLM for 3D generation and understanding[J]. ArXiv preprint arXiv: 2506.01853, 2025.
[16] LIPMAN Y, CHEN R T Q, BEN-HAMU H, et al. Flow matching for generative modeling[J]. ArXiv preprint arXiv: 2210.02747,2022.
[17] LIU B, ZHU Y F, GAO CH K, et al. LIBERO: Benchmarking knowledge transfer for lifelong robot learning[C]. Advances in Neural Information Processing Systems, 2023, 36: 44776-44791.
[18] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. ArXiv preprint arXiv: 1711.05101, 2017.
[19] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[J]. ArXiv preprint arXiv: 1608.03983,2016.
[20] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[J]. ArXiv preprint arXiv: 2106.09685,2021.
[21] CEN J, YU CH H, YUAN H J, et al. WorldVLA: Towards autoregressive action world model[J]. ArXiv preprint arXiv:2506.21539, 2025.
[22] CHI CH, XU ZH J, FENG S Y, et al. Diffusion policy: Visuomotor policy learning via action diffusion [J]. International Journal of Robotics Research, 2025,44(10/11):1684-1704.
[23] DASARI S, MEES O, ZHAO S, et al. The ingredients for robotic diffusion transformers [C]. 2025 IEEE International Conference on Robotics and Automation, 2025:15617-15625.

作者简介

覃宏伟,2024 年于东北大学获得学士学位,现为东北大学硕士研究生,主要研究方向为视觉-语言-动作模型。
E-mail:tanhw@mails.neu.edu.cn

Qin Hongwei received his B. Sc. degree from Northeastern University in 2024. He is currently pursuing his master's degree at Northeastern University. His main research interest includes vision-language-action (VLA) models.

姜杨(通信作者),2010 年于东北大学获得博士学位,现为东北大学副教授,主要研究方向为基于人工智能的工业机器视觉、基于工业知识增强的多模态大模型、工业机器人具身智能技术、机器人长期自主巡航等。
E-mail:jiangyang@mail.neu.edu.cn

Jiang Yang (Corresponding author) received his Ph. D. degree from Northeastern University in 2010. He is currently an associate professor at Northeastern University. His main research interests include AI-based industrial machine vision, industrial knowledge-enhanced multimodal large language models, embodied intelligence for industrial robots, and long-term autonomous navigation for robots.

白佳硕,2025 年于东北大学获得学士学位,现为东北大学硕士研究生,主要研究方向为视觉-语言-动作模型。
E-mail:2927845952@qq.com

Bai Jiashuo received his B. Sc. degree from Northeastern University in 2025. He is currently pursuing his master's degree at Northeastern University. His main research interest includes vision-language-action (VLA) models.

徐天尧,现为东北大学本科生,主要研究方向为多模态大语言模型。
E-mail:1993346550@qq.com

Xu Tianyao is currently an undergraduate student at Northeastern University. His main research interest includes multimodal large language models.

杨轩溥,现为东北大学本科生,主要研究方向为融合视觉感知。
E-mail:41662976@qq.com

Yang Xuanpu is currently an undergraduate student at Northeastern University. His main research interest includes integrated visual perception.