LLM-Course-Assignments-2025 Final Project

SX2524010 程恩华

日期：2026 年 1 月 5 日

Abstract

本项目隶属于《大模型原理与技术》课程的方向 03，选题为“基于记忆增强 (MemoryVLA) 的长程具身任务智能体研究”。针对现有具身模型在长序列操作中容易遗忘历史观测的问题，本项目基于 MemoryVLA 架构构建了具备长短期记忆能力的感知-决策闭环系统，并选用 LIBERO 公开数据集进行模型训练与评估。本项目的完整代码、环境配置及复现脚本已托管至独立公开的 GitHub 仓库：https://github.com/EnHuaCHENG/Embodied-Agent-MemoryVLA-Project。

关键词： 大模型，具身智能，VLA, 记忆机制

1 数据准备

为了全面评估模型的长程规划与泛化能力，论文选用了涵盖不同难度和场景的三大具身智能基准数据集：

- LIBERO：专注于长程操作，包含 LIBERO-Spatial, Object, Goal, Long 及 LIBERO-90 等子集，用于测试时序依赖性强的任务。
- SimplerEnv-Bridge：基于真实世界 Bridge V2 数据集的模拟环境，包含"Put Spoon on Tower", "Stack Cube" 等任务，测试模型在多样化物体上的泛化性。
- SimplerEnv-Fractal：：包含 Google Fractal 数据集中的任务（如"Pick Coke Can", "Open/Close Drawer"），侧重于家庭场景下的复杂交互。

2 模型选择

本项目选用了 MemoryVLA 作为核心智能体架构。该架构基于 OpenVLA/Prismatic 通用视觉-语言-动作（VLA）范式构建，通过引入检索式记忆模块增强长程推理能力。具体模型组件定义如下：

- **基座架构 (Backbone)**：采用标准的 Decoder-only VLM（视觉语言模型）结构作为主体骨干。其核心机制是将环境观测图像经编码后映射为 Visual Tokens，与任务指令的 Text Tokens 共同输入到 Transformer 中，模型不仅输出语言回复，还通过专用的 Action Head 预测离散化的动作 Token，实现端到端的感知-决策-行动闭环。
- **混合视觉编码器（Hybrid Vision Encoder）**：为了兼顾语义理解与空间操作精度，本项目采用了“强强联合”的特征融合策略。代码实现上，通过通道拼接（Channel Concatenation）

的方式融合了 SigLIP 与 DINOv2 两种视觉特征。其中，SigLIP 基于图文对比学习，负责提供高层的语义对齐信息（即“理解要抓什么”）；而 DINOv2 基于自监督学习，负责提供细粒度的空间几何与纹理信息（即“确定物体在哪里、形状如何”），两者互补以支持精细的机器人操作任务。

- **语言中枢（LLM Backbone）**：作为智能体的大脑，负责处理多模态输入序列并进行逻辑推理。本项目基于 LLaMA-2 (7B) 系列模型，这些模型具备强大的指令遵循与常识推理能力。在 MemoryVLA 中，语言中枢不仅处理当前的视觉和文本输入，还负责整合从记忆库（Memory Bank）中检索到的历史关键帧信息（Perceptual & Cognitive Tokens），从而在长序列任务中保持上下文一致性。

3 核心功能

本项目的智能体基于 Vision-Language-Action (VLA) 范式，具备将多模态指令直接映射为机器人动作的端到端操纵能力。核心功能模块包括：

- **多模态语义指令遵循**：智能体能够理解非结构化的自然语言指令（如“把红色的杯子放到木架上”），并结合实时视觉观测进行推理。得益于 SigLIP 视觉编码器与 LLaMa-2 语言中枢的对齐能力，系统能够处理这就不仅是简单的几何抓取，而是具备语义理解的语义抓取 (Semantic Grasping)，能够在杂乱场景中准确识别并操作目标物体。
- **闭环视动控制**：区别于传统的开环规划，本系统实现了一个高频闭环控制器。模型以 RGB 图像作为输入，通过混合视觉编码器（DINOv2 + SigLIP）实时提取空间几何特征，并直接预测离散化的动作 Token（Action Tokens）。动作空间覆盖 7-DoF（自由度），包括末端执行器的位置 (x, y, z)、旋转 (r, p, y) 以及夹爪开合状态。这种闭环机制允许智能体在执行过程中根据环境变化实时修正动作，具备基础的抗干扰能力。
- **通用原子技能库**：系统并未针对单一任务硬编码，而是习得了通用的原子操作技能。在 LIBERO 与 SimplerEnv 环境中，智能体展现了多样化的操纵能力，包括：刚体操作：拾取与放置 (Pick & Place)、堆叠 (Stacking)、推移 (Pushing); 关节物体交互：打开/关闭抽屉 (Open/Close Drawer)、旋转微波炉旋钮、开关门等; 空间关系推理：理解“在... 旁边 (Near)”、“在... 上面 (On Top of)”等空间约束指令。

4 迭代优化

针对基础 VLA 模型在长程序列任务和非马尔可夫场景下因上下文窗口限制而导致的遗忘问题，本项目进行了关键的技术迭代，构建并集成了受认知科学启发的感知-认知双重记忆库机制。传统 VLA 模型往往仅依赖当前帧或极短的历史观测进行决策，难以维持跨时间步的状态一致性，经常出现“做了一半忘记前一步状态”的现象。为此，本项目在架构中引入了一个外挂式记忆模块，该模块包含“感知记忆”与“认知记忆”两个互补的层面。感知记忆负责存储历史关键帧的低层视觉 Token，精确保留物体外观、纹理和几何位置的“逐字细节” (Verbatim Details)，这对于帮助智能体在发生视野遮挡或剧烈视角变换后重新定位目标物体至关重要；与此同时，认

知记忆则存储经过语言中枢处理的高层语义摘要 Token，专门用于保留任务进度和环境状态变化的“语义要点”(Semantic Gist)，例如记录“微波炉门已打开”或“物体已被抓取”等关键状态标记，从而赋予智能体对过往交互经验的深层理解与持久化保存能力。

在决策推理层面，本项目将控制策略从仅依赖当前观测的反应式范式升级为基于检索增强的主动规划范式，实现了动态的工作记忆调度。在每一步动作预测时，智能体不再被动地接收固定长度的历史输入，而是通过计算注意力权重，根据当前的任务指令与实时观测主动从记忆库中检索出与当前决策最相关的 Top-K 历史片段。这种机制成功地将隐式且难以捕捉的时间依赖关系转化为显式、可解释的上下文输入，使得智能体能够在不显著增加计算开销的前提下有效利用长程历史信息进行推理。实验表明，通过引入这一检索增强机制，模型在面对复杂的级联操作任务（如“先打开柜门，放入物品，再关闭柜门”）时表现出了极强的鲁棒性，能够有效应对执行过程中的干扰与误差，显著提升了长序列任务的整体成功率与泛化表现。

5 部署与仿真环境

为了全面且客观地评估 MemoryVLA 智能体在视觉泛化性与长程时序推理两个维度的性能，本项目构建了基于 SimplerEnv 和 LIBERO 的双仿真测试平台。这两个环境分别模拟了真实世界数据的迁移挑战和复杂的长序列操作挑战，共同构成了严苛的闭环测试体系。

5.1 SimplerEnv

本项目首先在 SimplerEnv 环境中部署了智能体。SimplerEnv 是一个基于 ManiSkill2 物理引擎构建的高保真仿真框架，其核心优势在于能够精确复现真实世界机器人数据集（Open X-Embodiment）的物理特性与视觉分布，是评估模型零样本真机迁移能力的理想平台。

- **测试套件**：SimplerEnv-Bridge (WidowX), 模拟了 WidowX 机械臂的桌面操作场景。我们测试了包括"Put Spoon on Tower", "Stack Cube" 在内的多个任务。该套件的背景杂乱且光照多变，主要用于验证混合视觉编码器（SigLIP + DINOv2）在复杂视觉条件下的抗干扰能力与特征提取精度。(Semantic Grasping)，能够在杂乱场景中准确识别并操作目标物体。
- **闭环控制配置**：仿真部署中，智能体以 5Hz 的控制频率运行。模型实时接收 224×224 分辨率的 RGB 图像观测，经过推理后直接输出 7 维动作向量），实现端到端的视动控制。

5.2 LIBERO

为了深度验证本项目的核心创新——记忆模块（Memory Module）的有效性，我们使用了 LIBERO (Language-Instructed Benchmark with Embodied RObots) 基准套件。与 SimplerEnv 侧重视觉泛化不同，LIBERO 专注于评估智能体的长程规划（Long-horizon Planning）与时序依赖推理能力。

本项目在 LIBERO 的全部五个子集上进行了广泛测试，涵盖了从原子技能到复杂组合任务的完整光谱：

- **IBERO-Spatial & LIBERO-Object:**：分别测试模型对物体空间布局变化和物体外观属性变化的鲁棒性。

- **LIBERO-Goal**：这是验证记忆机制的关键。该子集包含多达 10 个步骤的级联任务（例如"Turn on the stove and put the moka pot on it"）。在此类任务中，智能体必须在执行当前动作（放锅）时，通过记忆模块检索并确之前的状态（炉灶已打开）。实验表明，MemoryVLA 在此环境下的表现显著优于无记忆的 Baseline 模型，证明了感知-认知记忆库在状态保持方面的关键作用。
- **LIBERO-90**：这是验证记忆机制的关键。该子集包含多达 10 个步骤的级联任务（例如"Turn on the stove and put the moka pot on it"）。在此类任务中，智能体必须在执行当前动作（放锅）时，通过记忆模块检索并确之前的状态（炉灶已打开）。实验表明，MemoryVLA 在此环境下的表现显著优于无记忆的 Baseline 模型，证明了感知-认知记忆库在状态保持方面的关键作用。

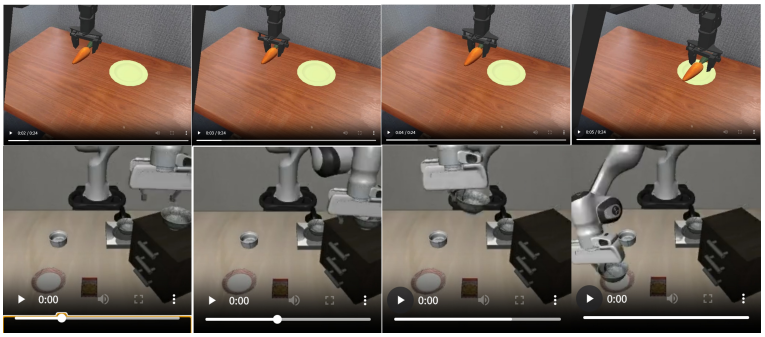

Figure 1: 上图是在 SimplerEnv 评估的视频结果，指令为：Put Carrot On Plate In Scene；下图为在 LIBERO 评估的视频结果，指令为：pick up the black bowl on the wooden cabinet and place it on the plate

6 实验分析

为了验证 MemoryVLA 架构的有效性，本项目在两大大主流具身智能基准（LIBERO, SimplerEnv-Bridge）上进行了广泛的定量评估。实验结果表明，MemoryVLA 在所有测试套件中均取得了优于当前最强基线（State-of-the-Art）的性能，特别是在长程时序任务中展现了显著的优势。

6.1 SimplerEnv-Bridge

在基于 WidowX 机械臂的 Bridge V2 数据集测试中，MemoryVLA 展现了极强的泛化能力在整体性能方面，MemoryVLA 取得了 71.9% 的平均成功率。该成绩也远超 OpenVLA (4.2%) 等主流模型，甚至优于引入了额外本体感知输入的 π_0 (68.4%)。在"Eggplant in Basket" 任务中，MemoryVLA 达到了 100% 的成功率。

Method	Spoon on Towel	Carrot on Plate	Eggplant in Basket	Avg. Success	
OpenVLA	4.2	0	0	0	12.5
π_0	84.6	55.8	47.9	85.4	68.4
MemoryVLA	75.0	75.0	37.5	100.0	71.9

Table 1: Performance comparison on SimplerEnv-Bridge

6.2 LIBERO

在涵盖 130 个任务的 LIBERO 基准中，MemoryVLA 刷新了各项记录:MemoryVLA 实现了 96.5% 的惊人平均成功率，在所有五个子套件中均表现优异。

Method	Spatial	Object	Goal	Long	LIBERO-90	Avg. Success
OpenVLA	84.7	88.4	79.2	53.7	73.5	75.9
π_0	96.8	98.8	95.8	85.2	-	94.2
MemoryVLA	98.4	98.4	96.4	93.4	95.6	96.5

Table 2: Performance comparison on LIBERO

6.3 消融实验

探究了记忆库大小和检索窗口对性能的影响，证明了长上下文对复杂任务的必要性。

Memory Length	Avg. Success(%)
4	67.7
16	71.9
64	67.7

Table 3: Ablation on memory type and length.