arXiv:2603.01465v1 [cs.RO] 2 Mar 2026

非马尔可 通过关键帧链式方法实现长时程机器人操作 g

陈一鹏[1] *谭文涛[1] *朱磊[1] ✉李凤玲[2] 李晶晶[3] 杨国立[4] 沈恒涛[1]

摘要

现有的视觉-语言-动作（VLA）模型往往难以泛化到长时程任务，因为它们严重依赖即时观测。尽管近期研究采用检索机制或扩展上下文窗口来处理程序性任务，但它们往往难以捕捉非马尔可夫依赖：在这种依赖中，最优动作仅取决于特定的过去状态，而非当前观测。为此，我们提出关键帧链式视觉-语言-动作模型（Keyframe-Chaining VLA），这是一种提取并关联关键历史帧以建模长时域依赖的框架。具体而言，我们提出一种自动关键帧选择器，用于学习判别式嵌入空间，从而有效识别不同的状态转换。为捕捉任务关键的信息，我们设计了一种进度感知查询机制，根据历史帧与当前执行阶段的时间相关性动态检索历史帧。这些已选关键帧作为交错的视觉词元集成到VLA中，使策略明确扎根于长时域时间上下文。最后，我们基于ManiSkill仿真器构建了一套包含四项非马尔可夫操作任务的基准，用于衡量任务成功率。实验结果表明，我们的方法取得了优异性能，能够有效应对具有长时域时间依赖的机器人操作任务。代码见
h
t
t
ps://github.com/cytoplastm/KC-VLA。

1. 引言

最先进的视觉-语言-动作（VLA）（Bjorck 等人，2025；Black 等人，2024；Team 等人，2024）模型展现出了卓越的控制能力，但仍受制于“密集观测”范式。依赖短暂且密集的历史窗口，隐含地假设存在马尔可夫性质，但在具有状态混叠特征的长时程任务中，这一假设便不再成立——此时，即时观测无法反映过去交互中的关键信息（Kaelbling 等人，1998）。虽然从理论上讲，扩展上下文窗口可以缓解这一问题，但二次注意力复杂度会带来计算瓶颈，使实时控制难以实现。现有方案，如潜在表示压缩或分层视觉语言模型规划，往往会牺牲精细的空间精度或推理速度。

为超越这些局限，我们提出了**关键帧链式视觉-语言-动作**框架，将时间抽象与动作生成解耦。我们没有让策略承担冗余视频流，而是引入轻量级的**关键帧选择模块（KSM）**，将连续观测提炼为由关键语义关键帧组成的稀疏语义历史。这样一来，策略便能像处理相邻帧一样轻松地推理相隔数千个时间步的事件，从而有效“压平”时间跨度，建立对任务结构的全局视角。

KSM采用了一种新颖的任务调制架构：通过在统一度量空间中利用特征级线性调制（FiLM），它能够根据任务语义动态调整检索标准，从而过滤运动噪声。随后，通过上下文感知提示策略，将这种稀疏且语义对齐的上下文注入流匹配策略中。该范式使智能体能够以极低的计算开销推理全局任务拓扑。

我们的贡献总结如下：我们提出了关键帧链式视觉-语言-动作模型，通过稀疏语义历史消除非马尔可夫歧义，绕过长上下文注意力的计算瓶颈。我们设计了任务调制 KSM，利用基于 FiLM 的查询生成，在统一的多任务度量空间中提取语义关键帧。我们建立了专为长时程语义推理定制的 ManiSkill 基准测试套件，弥补现有基准主要聚焦于马尔可夫设置的不足。大量实验表明，我们的方法在这些记忆密集型任务上取得了 92% 的成功率，显著优于基线方法（57%）。

*贡献相同 [1]同济大学 [2]悉尼科技大学 [3]电子科技大学 [4]大数据先进技术研究院。通信作者：朱磊 <leizhu0608@gmail.com>。

预印本版本。

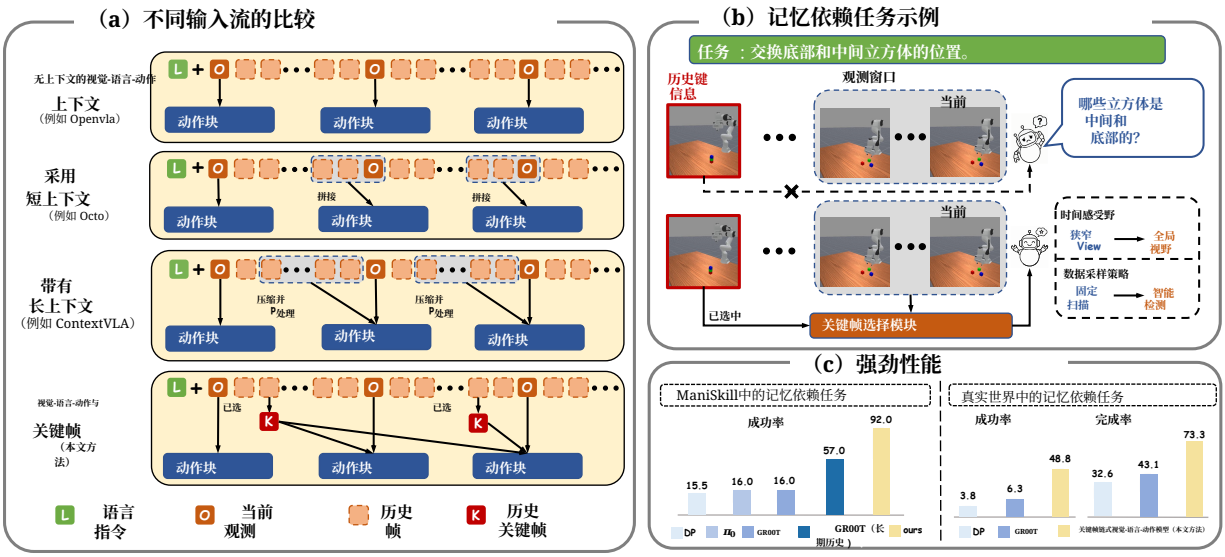

图 *1.* **(a)** 不同于依赖密集滑动窗口或上下文压缩的标准视觉-语言-动作模型（Kim 等人，2024；Team 等人，2024；Chi 等人，2025；Jang 等人，2025；Koo 等人，2025），我们的关键帧链式视觉-语言-动作模型会显式选择稀疏的语义关键帧，从而避免计算冗余。**(b)** (a) 中的其他三种范式受限于有限的时间感受野，无法捕捉远处的关键信号。相比之下，基于 KSM 的检索实现了全局时间感受野，使智能体能够访问长程历史信息，并有效解决状态混叠。**(c)** 我们的框架取得了最先进的性能，在 ManiSkill 中依赖记忆的任务以及真实世界的长时程部署中，均显著优于基线方法。

2. 相关工作

2.1. 视觉-语言-动作模型

RT-2（Zitkovich 等人，2023）和 OpenVLA（Kim 等人，2024）等开创性的视觉-语言-动作模型采用离散词元化来实现语义泛化，而 π_0（Black 等人，2024）和 GR00T（Bjorck 等人，2025）等近期最先进的模型则转向通过扩散或流匹配进行连续生成，以提升动作平滑度。尽管架构不断演进，主流输入范式仍局限于单张当前帧或密集的短期历史窗口。由于注意力机制的二次计算成本，这种设计不得不限制时间感受野，使得即便是先进的基础智能体也难免“目光短浅”，无法胜任需要大量长时程推理的任务。

2.2. 长上下文策略的记忆机制

为纳入历史上下文，近期研究探索了三种主要策略。首先，ContextVLA（Jang 等人，2025）和 HAMLET（Koo 等人，2025）等令牌压缩方法会将观测映射为紧凑的簇。然而，其压缩率往往不足以覆盖整个回合的历史信息，导致早期线索丢失。其次，MemER（Sridhar 等人，2025）等分层规划器利用视觉语言模型生成文本子目标。尽管逻辑上稳健，但依赖自然语言会带来一个瓶颈，限制细粒度空间信息的表达，并造成高昂的推理开销。第三，显式视觉保留方法侧重于流式记忆或跟踪。Mem- oryVLA（Shi 等人，2025）构建了一个记忆库，但递归融合会导致信息衰减。类似地，TraceVLA（Zheng 等人，2024）、SAM2Act（Fang 等人，2025）和 HiF-VLA（Lin 等人，2025）会显式跟踪物理动态，但缺乏抽象高层任务语义的能力。与这些方法不同，我们的框架采用轻量级选择模块，显式保留贯穿整个回合的离散语义关键帧序列，从而避免信息模糊和计算冗余。

2.3. 视频关键帧选择

计算机视觉领域已广泛探索长上下文视频理解（Zhang 等人，2023；Song 等人，2024），近期的研究工作如 AKS（Tang 等人，2025）和 Frame-voyager（Yu 等人，2024）采用关键帧选择来提高效率。然而，这些方法通常依赖计算量较大的多模态编码器或迭代式视觉语言模型查询，导致每帧计算成本和推理延迟较高。这些特性并不适用于闭环机器人控制，因为后者受到严格的实时延迟约束。此外，传统的视频摘要算法，如 DSNet（Zhu 等人，2020）、AC-SUM-GAN（Apostolidis 等人，2020）和 DR-DSN（Zhou

等人，2018）主要在离线场景中运行，依赖对完整序列进行全局优化。这与机器人操作的流式特性从根本上相悖。为实时控制而专门设计的方法采用轻量级度量学习模块，以极低的计算开销在线、流式地提取语义关键帧。

3. 预备知识

3.1. 语言条件控制策略

语言条件机器人策略通常经过训练，用于建模条件分布 $\pi(\mathbf{A}_t \mid o_t)$，其中$\mathbf{A}_t = [a_t, a_{t+1}, \ldots, a_{t+H-1}]$表示一个动作块，涵盖从当前时间 t开始的未来 H 个时间步。变量 o_t 表示机器人当前的感知观测。具体而言，观测表示为$o_t = [\mathbf{I}_t,\ l,\ q_t]$，其中 $\mathbf{I}_t = [I_t^1, I_t^2, \ldots, I_t^N]$ 由来自 N 相机视角的图像组成，l 是自然语言指令，q_t 表示本体感知输入（例如关节角度和夹爪状态）。

3.2. 非马尔可夫机器人操作

在标准的短时程任务中，当前观测 o_t通常包含足以推断潜在系统状态的充分统计量，从而满足马尔可夫性质。在此类场景中，将策略建模为 $\pi(\mathbf{A}_t \mid o_t)$即可实现成功执行。然而，复杂的操作任务通常会引入以部分可观测性为特征的**非马尔可夫**挑战。

例如，在烹饪任务中，判断是否已经加盐需要回忆过去的事件，因为汤的视觉状态可能保持不变。在这些情况下，即时观测 o_t 无法充分反映真实的环境状态（这一现象称为状态混叠），从而使瞬时策略 $\pi(\mathbf{A}_t \mid o_t)$失效。为解决此类任务，策略必须以交互历史为条件，形式化为 $\pi(\mathbf{A}_t \mid o_{0:t})$，从而利用时间上下文消解状态歧义。

3.3. VLA 输入模态：从密集到稀疏

视觉-语言-动作（VLA）模型作为骨干策略，旨在将多模态输入映射为连续机器人动作。输入模态从根本上决定了模型的时间推理能力。

传统输入流： 传统视觉-语言-动作模型通常采用单一的当前帧，或拼接一个较短的历史窗口，并将其建模为 $\pi(\mathbf{A}_t \mid o_t)$ 或 $\pi(\mathbf{A}_t \mid o_{t-k:t})$。由于注意力机制的计算复杂度呈二次增长，观测窗口大小 k受到严格限制。因此，这些模型往往无法应对非马尔可夫任务，因为它们不可避免地会丢弃密集滑动窗口之外发生的关键事件。

关键帧驱动输入流（本文方法）： 为克服时域限制，我们提出一种稀疏且具备语义感知的输入形式化。我们不再使用密集队列，而是维护一个由关键帧构成的记忆缓冲区——即在关键时间戳捕获的离散观测。我们的目标是确保历史关键帧与当前观测的组合能够捕获所有关键历史信息，从而近似完整历史：$\{o_{k_1}, o_{k_2}, \ldots, o_{k_n}, o_t\} \approx \{o_{0:t}\}$。相应地，VLA 输入流被重新形式化为$\pi(\mathbf{A}_t \mid o_{k_1}, o_{k_2}, \ldots, o_{k_n}, o_t)$。该范式使模型能够借助高效的记忆表示对长时域依赖进行推理。

4. 方法

4.1. 系统概述

我们提出了关键帧链式视觉-语言-动作框架。如图2所示，该架构通过两个协同模块实现稀疏输入范式：轻量统一的关键帧选择模块（KSM）和关键帧条件动作策略。具体而言，KSM 充当作用于视觉流$\{\mathbf{I}_0, \ldots, \mathbf{I}_t\}$的时间滤波器。当基于视觉线索检测到语义转变时，该模块会明确索引并保留相应的完整多模态观测 o_t（包括视觉和本体感觉），以形成离散的语义关键帧序列 $\mathcal{O}_{\text{key}} = \{o_{k_1}, \ldots, o_{k_n}\}$。随后，策略利用结构化输入格式组织多模态上下文，并基于检索到的这些关键帧和当前观测合成动作。这种解耦设计无需改变骨干网络的内部架构，即可实现高效的长时程推理。

4.2. 关键帧选择模块

我们提出了一种两阶段任务调制架构，用于提取语义关键帧。形式上，我们将每个长时程任务建模为由语义阶段构成的严格有序序列。我们将关键帧定义为标记各阶段完成的规范状态。因此，包含N 个阶段的任务由 N 个真实关键帧表征。检测到第 k个关键帧时，会触发状态转变，使系统从阶段 k推进到阶段 $k+1$。这些关键帧标注通过混合策略获取：对于语义关键帧伴随本体感觉状态明显变化的任务，我们利用这些信号自动提取关键帧；否则，我们采用人工标注以确保准确性。该模块学习一个统一潜在空间，使动态的任务条件查询能够实时识别这些关键状态转变。

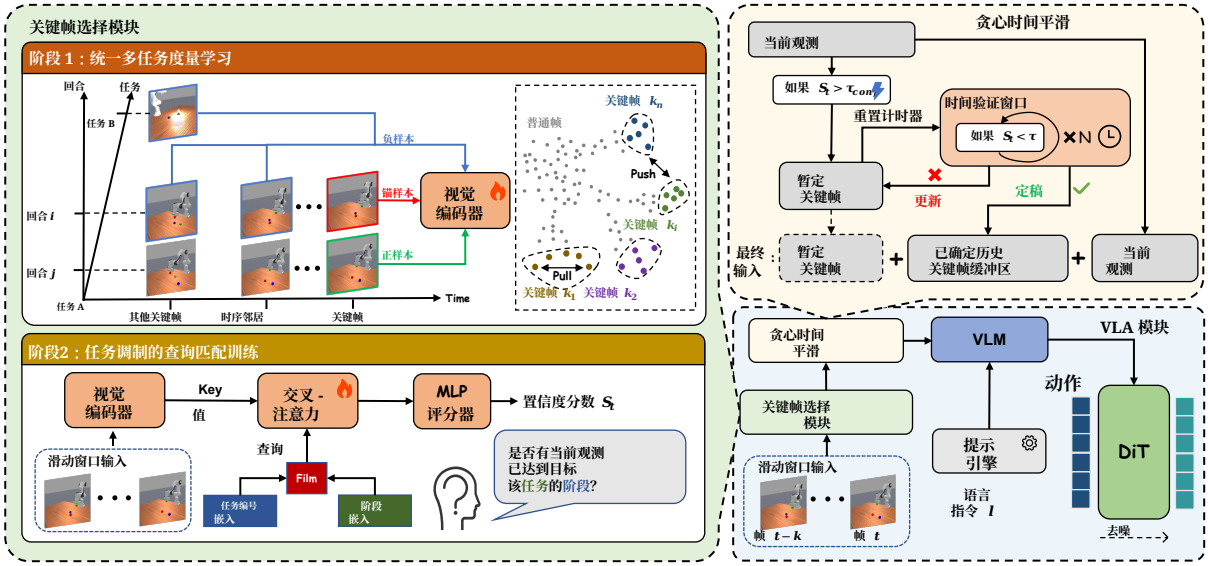

图 2。**关键帧链式视觉-语言-动作框架概览**。该架构通过两个协同工作的模块运行。**（左）关键帧选择模块：**采用两阶段设计，其中阶段I通过统一度量学习优化视觉编码器，而阶段二则使用任务调制查询网络计算特定阶段的置信度分数 s_t。**（右上）贪心时间平滑：**这是一种用于过滤检测抖动的稳定机制。该机制通过贪心更新暂定候选帧来捕捉最新的有效状态，并仅在通过时间验证窗口后才定稿关键帧。**（右下）视觉-语言-动作模块：**策略将检测到的关键帧所构成的稀疏语义历史与当前观测相结合，通过流匹配生成动作。

阶段I：统一多任务度量学习。我们采用 ResNet-18（He 等人，2016）E_θ 作为视觉编码器，将输入观测图像 I 映射为特征表示 $f(I) \in \mathbb{R}^d$。为确保特征边界具有判别性，我们通过三元组间隔损失（Schroff 等人，2015）对 E_θ 进行优化：

$$\mathcal{L}_{\text{triplet}} = \max(0, d(I_a, I_p) - d(I_a, I_n) + m). \quad (1)$$

具体而言，我们选择一个真实关键帧作为锚样本I_a，其由元组 (i, ψ, e) 表示，分别对应任务、阶段和回合。**正样本** I_p 取自同一阶段的另一个回合 (i, ψ, e')，其中 $e' \neq e$。这迫使编码器学习在不同回合中对每个阶段保持不变的表征。对于**负样本**I_n，我们采用平衡采样策略，以相同概率抽取以下三类负样本：

- **时间邻居** $\mathcal{N}_{\text{adj}}$：帧 $I_{t'}$ 位于同一回合内，且与锚样本的时间偏移满足$\delta_{\min} \leq |t' - t| \leq \delta_{\max}$，从而针对视觉相似的邻居进行细粒度区分。

- **任务内阶段负样本** $\mathcal{N}_{\text{phase}}$：从同一任务中 i 但属于不同阶段 $\psi' \neq \psi$的关键帧，确保编码器能够捕捉任务内独特的语义进展。

- **任务间负样本** $\mathcal{N}_{\text{任务}}$：从不同任务 $j \neq i$ 中采样的关键帧，确保全局任务可区分性。

这种三模态负样本采样迫使潜在空间同时遵循时间序列、阶段身份和任务边界。

阶段二：任务调制查询匹配训练。推理时，我们冻结在阶段I中训练的视觉编码器 E_θ，并采用轻量级查询网络来检测语义关键帧。首先，为了纳入局部时间上下文，我们将滑动窗口内的特征聚合为统一的表示 $\mathbf{H}_{\text{vis}}$：

$$\mathbf{H}_{\text{vis}} = \text{SelfAttn}\left(\text{Concat}[E_\theta(\mathbf{I}_{t-k:t})]\right). \quad (2)$$

接下来，我们合成一个动态逻辑查询$\mathbf{q}^{\text{logic}}$。具体而言，我们使用两组可学习嵌入来表示任务身份和通用阶段概念。给定自然语言指令 l，我们将其映射到相应的任务身份，以选取特定嵌入e_{task}。随后，我们通过生成器 g_ϕ将任务 ID 嵌入 e_{task} 投影为仿射参数 $[\gamma, \beta]$：

$$[\gamma, \beta] = g_\phi(e_{\text{task}}). \quad (3)$$

这些参数随后通过 FiLM 调制共享的阶段嵌入e_{phase}（Perez 等人，2018），以编码任务特定语义：

$$\mathbf{q}_{\text{logic}} = \gamma \odot e_{\text{phase}} + \beta. \quad (4)$$

最后，查询 $\mathbf{q}^{\text{logic}}$ 通过交叉注意力从 $\mathbf{H}_{\text{vis}}$中检索相关线索。我们计算匹配分数 s_t：

$$s_t = \text{MLP}(\text{CrossAttn}(Q = \mathbf{q}_{\text{logic}}, K{=}V{=}\mathbf{H}_{\text{vis}})). \quad (5)$$

当匹配分数 s_t 超过预定义的置信度阈值 τ_{conf}时，系统会识别出一个里程碑完成。因此，当前观测 o_t 会被检索并追加到稀疏历史缓冲区中，同时更新阶段指针（$id_{\text{phase}} \leftarrow id_{\text{phase}} + 1$）。

平衡训练协议。为缓解关键帧稀疏问题并增强鲁棒性，我们采用阶段引导的平衡采样策略，生成带有二元目标标签 $y \in \{0, 1\}$ 的训练对：

- **正样本**（$y = 1.0$）：采样真实关键帧及其紧邻时间邻居（偏移量为 $\delta \in \{-1, 0, 1\}$）对应的帧，以适应轻微的标注抖动。
- **轨迹内负样本**（$y = 0.0$）：从轨迹片段 $[k_{i-1}, k_i]$ 内的 M 个等距区间随机采样帧，以抑制误触发。
- **阶段不匹配负样本**（$y = 0.0$）：将有效里程碑图像 o_{k_i} 与来自后续阶段的查询 $\mathbf{q}$（id_{phase} +**1**）配对，以强制遵循严格的顺序逻辑。

推理与贪心平滑。为确保机器人实时执行过程中能够抵抗分数波动并稳健地检测里程碑，我们采用了带重置功能的贪心时间平滑机制。当匹配分数满足 $s_t > \tau_{\text{conf}}$ 时，系统会将当前观测缓存为暂定关键帧，并启动时间验证窗口。该窗口用于稳定性测试：如果窗口内后续任一帧再次超过 τ_{conf}，则该帧会成为更强的候选帧，立即取代当前暂定关键帧并重置验证计时器。该机制确保系统贪心地锁定最新的有效帧，从而捕捉子任务完成的确切状态。只有当分数在整个窗口期间持续低于 τ_{conf} 时，候选帧才会被提交到最终历史缓冲区 $\mathcal{O}_{\text{key}}$；此时阶段指针随之递增。

4.3. 关键帧链式策略架构

我们采用最先进的通用策略 GR00T-N1.5 作为骨干网络，以验证所提出框架的有效性。其基于Transformer的架构天然支持可变长度令牌序列，因此非常适合作为整合非均匀稀疏历史的基础。为了有效利用提取出的关键帧，我们重新设计了 GR00T-N1.5 的感知流，将冗余的密集观测改为稀疏语义历史。具体而言，我们构建了一个非均匀输入序列$\mathcal{I} = \{o_{k_1}, \ldots, o_{k_n}, o_t\}$，将检索到的历史关键帧与当前视觉观测拼接起来。为了对这一时间序列进行精确推理，我们采用结构化的文本系统提示，明确强调时间关系，并引导模型分析任务进展。最终，这一语义对齐的序列 $\mathcal{I}$被注入流匹配动作头（Lipman等人，2022）作为条件。通过将策略建立在任务的拓扑进展之上，该架构解决了单帧观测固有的状态混叠，使智能体能够超越马尔可夫限制，并生成与完整逻辑链一致的动作。

5. 实验

本节在高保真仿真环境和真实环境中开展了一系列全面的实验，系统评估我们提出的关键帧链式视觉-语言-动作框架的有效性。实验设计聚焦于以下三个维度：

1. **稀疏语义历史的有效性：**我们提出的稀疏历史机制能否有效缓解非马尔可夫任务中的状态混淆，并优于依赖密集观测窗口或朴素采样启发式方法的基线方法?
2. **KSM 的准确性与必要性：**关键帧选择模块（KSM）检测语义里程碑的诊断精确率如何？其内部组件对于构建判别性特征空间是否不可或缺？ 3. **物理部署与鲁棒性：**对于直接在真实环境中训练的策略，稀疏关键帧链式机制在多大程度上能够在固有的环境噪声和感知干扰下维持稳定的长时程控制?

5.1. 仿真：记忆依赖基准

弥合仿真评估差距。推动基于记忆的机器人学习的一大关键瓶颈，是缺乏合适的评估平台。现有基准，如 LIBERO（Liu 等人，2023）、RLBench（James 等人，2020）和 CALVIN（Mees 等人，2022），通常侧重于程序复杂性，但整体上仍以马尔可夫特性为主——也就是说，仅凭当前观测就能推断出最优动作（例如“打开看得见的抽屉”）。当前数据集中的这种“马尔可夫偏差”未能对短视策略予以惩罚，从而掩盖了它们处理状态混叠的能力缺陷。

为弥补这一不足，我们开发并开源了一个专门的**长时程记忆依赖基准**，该基准基于高保真 ManiSkill（Gu 等人，2023）仿真器构建。有关环境初始化、成功指标和真实关键帧定义的详细规范见附录A。与以往的测试套件不同，我们的基准引入了严格的非马尔可夫约束，其中关键任务信息（例如，过去的交互历史）会被严格隐藏，不纳入当前观测。

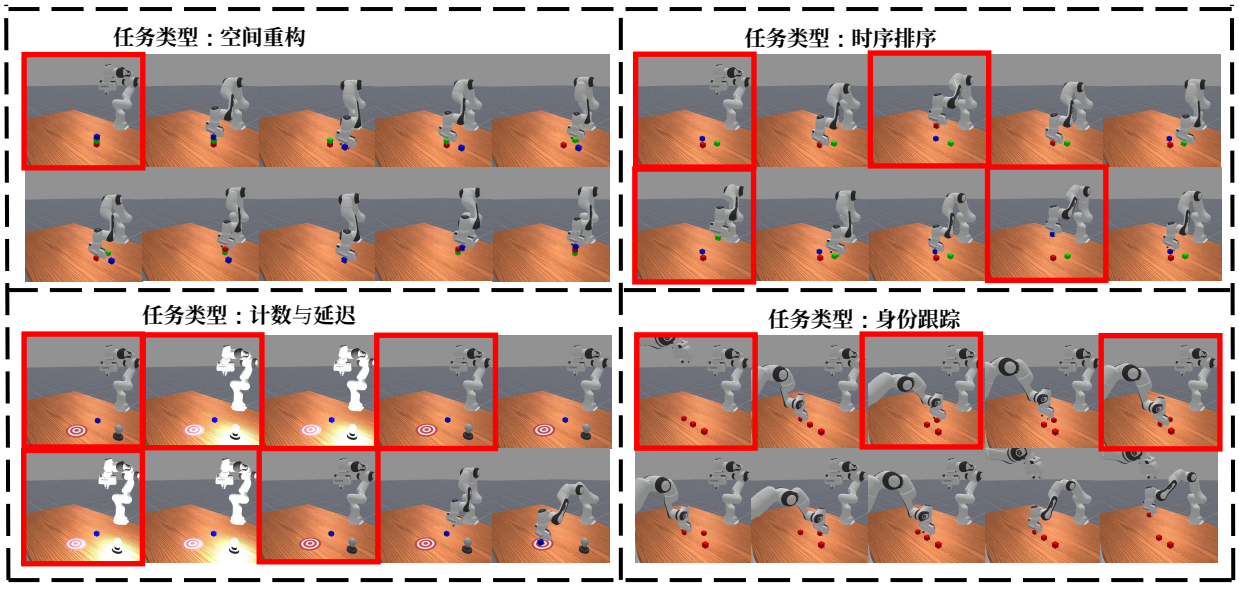

图 *3.* **四个定制的非马尔可夫 ManiSkill 任务概览**。红色边界框标示了训练中使用的真实关键帧。各任务关键帧事件的详细定义请参阅**附录**。

该测试套件为学界提供了一个严格的测试平台，用于评估智能体维持持久内部状态和进行时间推理的能力。

任务协议。 我们设计了四项任务（见图 3），它们需要不同类型的记忆依赖：

- **空间重构**：智能体必须拆解由三个随机排序方块组成的垂直堆叠，并按置换后的顺序重新构建它们。**理由**：主要难点在于“拆解—重建”循环。一旦拆开，关于初始相对顺序的视觉信息便永久丢失。智能体必须完全依赖对初始配置帧的记忆，才能执行正确的重新堆叠顺序。

- **时序排序**：机器人必须严格按照 $Red \rightarrow Green \rightarrow Blue$的顺序，对三个彩色方块执行“抓取—抬起—复位”循环。**理由**：该任务在任务进展方面引入了严重的状态混叠。例如，仅凭悬停的夹爪，无法从静态观测中判断哪些方块已被操作。智能体必须通过检索到的关键帧核验子目标的历史完成情况，以区分有效进展与顺序违规。

- **计数与延迟**：信号灯以随机间隔闪烁两次。智能体必须对这些脉冲进行计数，并且只有在第二次闪光后才能推目标。**理由**：这用于评估在静态间隔期间维持内部状态的能力。由于闪烁间隔期间的场景与初始状态无法区分，马尔可夫策略无法在没有持久历史上下文的情况下确定当前计数。

- **身份跟踪**：三个外观完全相同的红色方块整齐排列，一只辅助机械臂在其中两个方块之间进行了快速交换。智能体的任务是拾取最初位于中央的那个特定方块。**理由**：由于这些物体具有完全相同的视觉特征，目标的身份完全由其历史决定。模型必须通过检索记录交换事件的特定历史帧来进行物体恒常性推理，而没有精确的时间回忆能力，这一能力从根本上无法实现。

基线方法。我们将关键帧链式视觉-语言-动作模型与具有代表性的最先进机器人策略子集进行比较，在各种时间配置下对其进行评估，以建立稳健的基准：

- **扩散策略（DP）**（Chiet 等人，2025）：一种领先的行为克隆方法，使用条件扩散过程对动作分布进行建模。我们采用标准实现，并使用较短的观测时域（$T_{obs} = 2$），以代表强短期策略。

- π_0 **（流匹配）**（Black 等人，2024）：一种最先进的通用型视觉-语言-动作策略，采用流匹配实现高保真动作生成。与扩散策略类似，它主要依赖即时视觉上下文进行决策。

- **GR00T-N1.5**(Bjorck 等人，2025)：一种大规模通用视觉-语言-动作模型。为了系统评估时间感受野的影响，我们对 GR00T 进行全面分析，

*表1。*非马尔可夫任务成功率（%）的定量比较。关键帧链式视觉-语言-动作模型（本文方法）取得了稳健且优异的表现。

模型 / 配置	采样	N_h	I	空间	时间	身份	计数	平均值
π_0 (Black 等人,2024)	密集	0	-	2.0	0.0	0.0	60.0	15.5
扩散策略 (Chi 等人, 2025)		1	1	22.0	10.0	0.0	30.0	15.5
GR00T-N1.5（Bjorck 等人，2025）（无历史）		0	-	20.0	0.0	28.0	16.0	16.0
GR00T-N1.5（短期）	密集	1	1	8.0	16.0	30.0	4.0	14.5
		2	1	18.0	52.0	12.0	0.0	20.0
		3	1	14.0	54.0	0.0	40.0	27.0
GR00T-N1.5（长期）	固定步幅	3	5	20.0	80.0	32.0	30.0	40.5
		3	10	18.0	94.0	26.0	60.0	49.5
		3	20	18.0	94.0	22.0	74.0	52.0
		3	30	6.0	38.0	44.0	14.0	25.5
		3	40	60.0	84.0	84.0	0.0	57.0
		3	50	16.0	66.0	52.0	60.0	48.5
		3	60	26.0	70.0	78.0	2.0	44.0
		3	70	24.0	40.0	32.0	2.0	25.0
		3	80	0.0	22.0	8.0	22.0	13.0
		3	90	28.0	52.0	50.0	8.0	35.0
		3	100	2.0	14.0	4.0	4.0	6.0
关键帧链式视觉-语言-动作模型（本文方法）	**关键帧**	-	-	**70.0**	**98.0**	**100.0**	**100.0**	**92.0**

分析 GR00T 在三种不同记忆机制下的表现，这些机制由历史长度（N_h）和采样间隔（I）定义：（i）无历史（$N_h = 0$）：纯粹的马尔可夫基线，仅依赖当前帧。（ii）短期历史（$N_h \in \{1, \ldots, 3\}$，$I = 1$）：密集的观测窗口，用于捕捉局部时间线索。（iii）长期历史（$N_h = 3$，$I \in \{5, \ldots, 100\}$）：采用固定步幅采样方法，以牺牲时间分辨率为代价扩大时间跨度覆盖。

定量分析。如表 1所示，关键帧链式视觉-语言-动作模型展现出卓越性能，平均成功率达到 **92.0%**，以显著间隔超越最强基线（57.0%）。进一步考察基线行为可以发现，标准记忆机制存在根本性局限。具体而言，依赖短时、密集观测窗口（$N_h \leq 3$、$I = 1$）的策略在所有非马尔可夫任务中均持续失败，平均成功率低于30%。这一普遍失败表明，仅仅堆叠即时观测不足以支持长时程推理，因为任务关键线索——例如空间重构中的初始方块配置——往往会退至这一狭窄时间跨度之外。

此外，尽管通过固定步长采样（$I > 1$）扩展感受野试图缓解这种短视问题，但这会在**时间跨度**和**分辨率**之间引入固有的权衡。不同任务的性能差异清楚地证明了这一冲突：较粗的采样率（例如 $I = 40$）能够成功捕捉空间重构 所需的远端初始状态（成功率为 60%），但在计数 任务中却完全失败（0%），原因是信号混叠导致帧间快速闪烁的灯光被跳过。相反，较小的步长（$I = 20$）能够解析高频计数信号（74%），但无法涵盖空间推理所需的历史信息（18%）。因此，没有任何单一的固定采样率能够泛化到时间动态特性各异的任务。

相比之下，我们的方法通过从固定间隔扫描转向事件驱动检索，规避了这种超参数权衡。通过明确选择具有语义重要性的帧，无论事件持续时间或发生时机如何，模型都能实现稳定的高性能（例如，在身份和计数任务中达到100%）。这一结果表明，长时程操作的瓶颈并不只是记忆容量，而在于检索到的上下文的语义相关性，从而验证了我们的假设：稀疏且具有区分性的历史表示，比密集且含噪的窗口更为有效。

实现细节。我们整理了一个包含 400 个回合的多任务数据集，每个非马尔可夫任务均衡分配 100 个回合。策略架构将一个包含 30 亿参数的视觉语言模型骨干网络用于语义推理，并集成一个包含 5 亿参数的扩散变换器头，以实现精准的动作生成。训练在单张 NVIDIA L40 GPU 上进行，共训练 50,000 步，批量大小为 16，采用 AdamW优化器，学习率恒定为 1×10^{-4}。

5.2. KSM评估与消融实验

诊断设置。我们整理了一个包含 400 个回合（每项任务 100 个）的集中式数据集，用于评估关键帧选择模块（KSM）。我们按 80/20 的比例划分数据，为 KSM 任务调制查询网络构建训练集，并

构建留出测试集，以验证其对未见执行轨迹的泛化能力。

度量定义。 应用时间贪心平滑机制，我们将 5帧窗口内的原始触发索引聚类，并使用簇中位数作为代表性时间戳。如果簇中位数与真实值（GT）之间的时间偏移在±10 帧的容差范围内，则将预测分类为真正例（TP）。假正例（FP）表示冗余或未对齐的触发，而假负例（FN）表示遗漏的关键帧。基于这些分类，我们报告精确率、召回率和 F1分数。此外，我们使用假负例率和假正例率分析错误模式。

表2。 KSM范式的比较诊断评估。各项度量在四项任务上取平均值，并以百分比表示。所有度量均以百分比（%）表示。各任务的详细结果见附录 B。

KSM训练范式	P (%) ↑	R (%) ↑	F1 (%) ↑	FPR (%) ↓	FNR (%) ↓
无度量预训练	90.7	81.7	85.1	9.3	18.3
联合端到端	93.6	79.6	83.4	6.4	20.4
我们的方法（两阶段）	**97.5**	**97.5**	**97.5**	**2.5**	**2.5**

训练范式分析。表2 突显了解耦训练策略的有效性，其 F1分数 达到近乎完美的 97.5。至关重要的是，端到端联合训练范式表现出较高的假负例率（20.4）。这种性能下降在计数任务中最为明显（召回率=40.0），这表明若没有三元组损失所约束的判别式嵌入空间，模型便无法将瞬态语义事件与静态背景区分开来。通过严格要求先进行度量学习，再进行查询训练，我们的两阶段方法确保视觉-语言-动作模型获得高保真的稀疏历史 $\mathcal{O}_{\text{key}}$。

表3。 **提示词优化消融实验。**比较基于上下文的提示词与通用指令。

配置	空间	时间身份		计数	平均值
不进行提示词优化	56.0	96.0	100.0	100.0	88.0
我们的方法（上下文感知）	**70.0**	**98.0**	**100.0**	**100.0**	**92.0**

提示词优化的必要性。我们通过将优化后的提示词与通用指令进行比较，分离出上下文感知提示的影响。表3 显示出对任务的敏感性：空间重构在不进行优化时性能显著下降（70% → 56%）。为此，我们设计了一种结构化系统提示词，明确地
将输入定义为按时间顺序排列的多视角图像对序列。具体的提示词表述见附录 C。

5.3. 真实世界实验

为验证我们的方法在实际场景中的有效性，我们使用一台 AgileX Piper 机械臂。我们重点评估关键帧链式视觉-语言-动作模型能否在物理环境中持续进行长时程推理。为进行策略适应，我们为每项任务收集了 **50 条专家示范**，通过傀儡臂装置进行远程操作。附录 D中提供了这些任务的执行过程可视化结果。

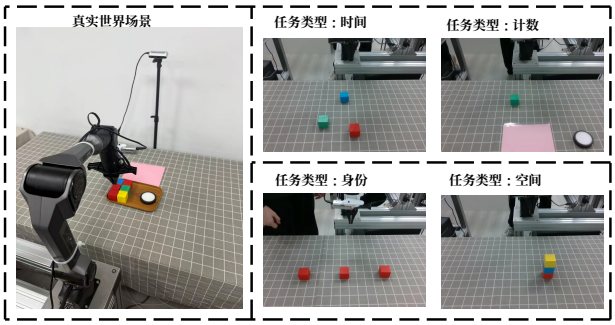

图 *4.* 真实世界实验设置。

表 4. **真实世界定量评估（每项任务进行 20 次试验）。**结果以**成功率**/完成率（%）报告。完成率表示已完成子阶段的百分比（阶段定义见附录）。我们的方法在多阶段推理任务中展现出显著优势。

模型	空间 (4个阶段)	时间计数 (3个阶段)	(2个阶段)	身份 (2个阶段)	平均值
DP	0.0 / 33.8	0.0 / 16.7	5.0 / 35.0	10.0 / 45.0	3.8 / 32.6
GR00T	0.0 / 47.5	0.0 / 30.0	10.0 / 45.0	15.0 / 50.0	6.3 / 43.1
Ours	**10.0 / 58.8**				

定量分析。 在 20 次试验中进行评估后，我们的方法展现出显著的性能优势。在评估的各项任务中，我们的方法在成功率（SR）和完成率（CR）方面均取得了稳定提升。与常因早期阶段错误而受挫的基线方法不同，我们的策略在较长时间范围内保持语义一致性，显著提高了成功率；即使任务仅部分完成，也展现出更优的阶段性进展。总体而言，这一全面优势验证了该框架在物理环境中的几何与时间定位能力得到提升。

6. 结论

我们提出了**关键帧链式视觉-语言-动作模型**，通过稀疏语义历史解决了非马尔可夫歧义。我们的任务调制 KSM 提取事件驱动关键帧，以高效建立长时域依赖，因而在 ManiSkill 基准上取得了 92.0% 的成功率，并在真实世界中表现稳健。**局限性与未来工作。** 存储原始视觉关键帧会造成像素空间

冗余，且存储规模呈线性增长。未来工作将探索把历史压缩为紧凑的潜在抽象或离散语义标记，以提炼任务相关信号，同时研究动态记忆更新机制。

参考文献

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V. 和 Patras, I. AC-SUM-GAN：连接行动者-评论家与生成对抗网络，实现无监督视频摘要。*IEEE* 视频技术电路与系统汇刊，31(8):3278–3292，2020。

Bjorck, J., Casta˜neda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S. 等。Gr00t n1：面向通用型人形机器人的开放基础模型。*arXiv* 预印本 *arXiv:2503.14734*，2025。

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B. 等。 π_0：用于通用机器人控制的视觉-语言-动作流模型。*arXiv* 预印本 *arXiv:2410.24164*，2024。

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burch- fiel, B., Tedrake, R., and Song, S. 扩散策略：通过动作扩散进行视觉运动策略学习。《国际机器人研究期刊》，44(10-11)：1684–1704，2025。

Fang, H., Grotz, M., Pumacay, W., Wang, Y. R., Fox, D., Krishna, R., and Duan, J. SAM2Act：将视觉基础模型与记忆架构相结合，用于机器人操作。*arXiv* 预印本 *arXiv:2501.18564*，2025。

Gu, J., Xiang, F., Li, X., Ling, Z., Liu, X., Mu, T., Tang, Y., Tao, S., Wei, X., Yao, Y., et al. Maniskill2：用于可泛化操作技能的统一基准。*arXiv* 预印本 *arXiv:2302.04659*，2023。

He, K., Zhang, X., Ren, S., and Sun, J. 用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。

James, S., Ma, Z., Arrojo, D.R., and Davison, A.J. RLBench：机器人学习基准与学习环境。*IEEE*机器人与自动化快报，5(2):3019–3026，2020年。

Jang, H., Yu, S., Kwon, H., Jeon, H., Seo, Y., and Shin, J. ContextVLA：采用摊销多帧上下文的视觉-语言-动作模型。*arXiv*预印本 *arXiv:2510.04246*，2025年。

Kaelbling, L. P.、Littman, M. L. 和 Cassandra, A. R. 在部分可观测随机领域中的规划与行动。人工智能，101(1-2):99–134，1998。

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。Koo, M.、Choi, D.、Kim, T.、Lee, K.、Kim, C.、Seo, Y. 和 Shin, J. Hamlet：将视觉-语言-动作模型转换为历史感知策略。*arXiv* 预印本 *arXiv:2510.00695*，2025。

Lin, M., Ding, P., Wang, S., Zhuang, Z., Liu, Y.,Tong, X., Song, W., Lyu, S., Huang, S., and Wang, D. HiF-VLA：通过运动表征实现视觉-语言-动作模型的后见、洞察与前瞻。*arXiv*预印本 *arXiv:2512.09928*，2025年。

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., 和 Le, M. 用于生成建模的流匹配。*arXiv* 预印本 *arXiv:2210.02747*，2022。

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., 和 Stone, P. LIBERO：面向终身机器人学习的知识迁移基准。神经信息处理系统进展，36:44776–44791，2023。

Mees, O., Hermann, L., Rosete-Beas, E., 和 Burgard, W. CALVIN：面向长时域机器人操作任务的语言条件策略学习基准。*IEEE*机器人与自动化快报，7(3):7327–7334，2022。

Perez, E., Strub, F., De Vries, H., Dumoulin, V., 和 Courville, A. FiLM：通过通用条件层进行视觉推理。载于*AAAI*人工智能会议论文集，第32卷，2018。

Schroff, F., Kalenichenko, D., 和 Philbin, J. Facenet：用于人脸识别和聚类的统一嵌入。载于*IEEE*计算机视觉与模式识别会议论文集，第 815–823 页，2015。

Shi, H., Xie, B., Liu, Y., Sun, L., Liu, F., Wang, T., Zhou, E., Fan, H., Zhang, X., 和 Huang, G. Memo-ryvla：用于机器人操作的视觉-语言-动作模型中的感知-认知记忆。*arXiv* 预印本 *arXiv:2508.19236*，2025。

Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y. 等。Moviechat：从密集令牌到稀疏记忆，以理解长视频。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18221–18232 页，2024。

Sridhar, A.、Pan, J.、Sharma, S. 与 Finn, C. MemER：通过经验检索扩展机器人控制的记忆。*arXiv* 预印本 *arXiv:2510.20328*，2025。

Tang, X.、Qiu, J.、Xie, L.、Tian, Y.、Jiao, J. 与 Ye, Q. 面向长视频理解的自适应关键帧采样。见计算机视觉与模式识别会议论文集，第29118–29128页，2025。

Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Krei man, T., Xu, C. 等。Octo：一种开源通用机器人策略。*arXiv* 预印本 *arXiv:2405.12213*，2024。Yu, S., Ji n, C., Wang, H., Chen, Z., Jin, S., Zuo, Z., Xu, X., Sun, Z., Zhang, B., Wu, J. 等。Frame-voyager：学习为视频大语言模型查询帧。*arXiv* 预印本 *arXiv:2410.03226*，2024。Zhang, H., Li, X. 和 Bing, L. Video-llama：一种用于视频理解的指令微调音视频语言模型。*arXiv* 预印本 *arXiv:2306.02858*，2023。Zheng, R., Liang, Y., Huang, S., Gao, J., Daum´e III, H., Kolobov, A., Huang, F. 和 Yang, J. Tracevla：视觉轨迹提示增强通用机器人策略的时空感知能力。*arXiv* 预印本 *arXiv:2412.10345*，2024。Zhou, K., Qiao, Y. 和 Xiang, T. 通过多样性-代表性奖励进行无监督视频摘要的深度强化学习。载于*AAAI*人工智能会议论文集，第32卷，2018。Zhu, W., Lu, J., Li, J. 和 Zhou, J. DSNet：一种用于视频摘要的灵活检测到摘要网络。*IEEE*图像处理汇刊，30：948–962，2020。Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A. 等。RT-2：视觉-语言-动作模型将网络知识迁移到机器人控制。载于机器人学习会议，第2165–2183页。PMLR，2023。

A. 任务规格与预言机设计

在本节中，我们详细介绍 e实现逻辑、成功标准以及真值（预言机）关键帧选择heuristi
cs，针对四个自定义的非Markovian任务，在ManiSkill中设计。

A.1. 时序排序（PickPlaceThreeTimes）

指令： “首先，拿起红色方块并将其放回桌面。接着，对绿色方块执行相同操作。最后，处理蓝色方块。”

• **环境设置：** 三个彩色方块（红色、绿色、蓝色）以随机姿态初始化在桌面上。放置目标位置与它们的初始坐标重合。

• **成功标准：** 当且仅当满足以下条件时，任务才算成功：

1. 顺序抬升： 检测到红色、绿色和蓝色方块严格按 $Red \rightarrow Green \rightarrow Blue$所示顺序超过高度阈值 h_{lift}。
2. 精确放置： 任务完成后，每个方块的最终位置与其初始位置之间的欧氏距离均在容差阈值 ϵ 以内（即返回原位）。

• **预言机关键帧定义（4 帧）：** 为训练 KSM，我们根据物体的 z 轴轨迹选择以下真实关键帧：

1. **帧 0：** 包含目标构型的初始观测。2. **帧** t_R**：** 红色方块达到最大高度（z_{max}）时的时间戳。3. **帧** t_G**：** 绿色方块达到最大高度时的时间戳。4. **帧** t_B**：** 蓝色方块达到最大高度时的时间戳。

A.2. 计数与延迟（PushCubeWithSignal）

指令： “等待信号灯闪烁两次，然后将方块推向目标。”

• **环境设置：** 方块随机初始化在一个位置。信号灯经历如下状态转移序列：$Off \rightarrow On \rightarrow Off \rightarrow On \rightarrow Off$。前四种状态的持续时间随机化，以防止时间过拟合。

• **成功标准：**

1. 目标达成： 将方块推至目标区域。2. 约束满足： 在信号灯进入最终的 Off 状态之前（即第二次闪光结束之前），方块的位移必须保持为零。任何在此触发条件达成前发生的移动都会导致立即失败。

• **预言机关键帧定义（5 帧）：** 在信号状态转变（信号边沿）的准确时刻选取关键帧：

1. **帧 0：** 初始状态。 2. **帧** t_1**：** 上升沿（第 1 次闪烁开始：$Off \rightarrow On$）。 3. **帧** t_2**：** 下降沿（第 1 次闪烁结束：$On \rightarrow Off$）。 4. **帧** t_3**：** 上升沿（第 2 次闪烁开始：$Off \rightarrow On$）。 5. **帧** t_4**：** 下降沿（第 2 次闪烁结束：$On \rightarrow Off$）。

A.3. 空间重构（SwapPosition）

指令： “交换底部和中间立方体的位置。”

• **环境设置：** 三个彩色方块（红色、绿色、蓝色）以单个垂直堆叠的形式初始化。堆叠顺序和基座位置均为随机设置。

- **成功标准：** 该任务检查方块的相对垂直顺序（z 坐标）：

  1. 最初位于顶部的方块必须保持在最高的 z 位置。2. 最初位于底部的方块必须处于中间的 z 位置。3. 最初位于中间 的方块必须处于最低的 z 位置。

- **预言机关键帧定义（1 帧）：** 关键帧记录了记住初始配置所需的关键拆解步骤:

  1. **帧 0：** 初始堆叠状态（对于识别原始顺序至关重要）。

A.4. 身份跟踪（教师机械臂洗牌）

指令：“立方体交换后，拿起最初位于中间的那个方块。”

- **环境设置：** 三个外观完全相同的红色方块水平排列。自动化的“教师机械臂”会对随机选出的两个方块（分别记为立方体 A 和立方体 B）进行随机洗牌操作。洗牌步骤如下：（1）将 A 移至临时缓冲区。（2）将 B 移至 A 的原始位置。（3）将 A 移至 B 的原始位置。
- **成功标准：** 教师机械臂完成洗牌步骤并收回后，智能体必须抓取并抬起在 $t = 0$ 时对应于几何中心的特定物体 ID，使其超过高度阈值 h_{lift}。
- **预言机关键帧定义（3 帧）：** 关键帧聚焦于交换轨迹，以确定物体身份:

  1. **帧 0：** 确定“中间”身份的初始状态。2. **帧 t_A：** 教师机械臂将立方体 A 放入临时缓冲区的瞬间。
  3. **帧 t_B：** 教师机械臂抓住立方体 B 的瞬间（表示第二次位移开始）。

B. 按任务划分的 KSM 诊断分析

在第 5.2 节中，我们报告了关键帧选择模块（KSM）的汇总诊断指标。这里，我们按四项任务细分展示性能，并将我们提出的两阶段方法与“端到端”和“无度量预训练”基线方法进行比较。

如表5所示，由于光信号（快速闪烁）具有瞬态特性，基线方法在**计数**任务中面临的挑战最大。端到端模型在该任务上的召回率大幅下降至 40.0%，这证实了若缺少 StageI 提供的显式度量约束，模型便难以将语义上的“信号”与背景噪声区分开来。相比之下，我们的方法在所有任务中都保持了近乎完美的检测率（100.0%），仅在涉及破坏-重建循环中复杂遮挡的空间重构任务上略有下降（90.0%）。

表5. **按任务划分的 KSM 详细诊断指标（%）**。我们比较了三种训练范式下的精确率（P）、召回率（R）和 F1分数。我们的方法在所有类别中始终达到最优或接近最优的性能。

Task	方法	精确率	召回率	F1分数
时序排序	无度量预训练	80.3	81.7	81.0
	联合端到端	93.3	93.3	93.3
	我们的方法（两阶段）	**100.0**	**100.0**	**100.0**
计数与延迟	无度量预训练	**100.0**	62.5	76.9
	联合端到端	**100.0**	40.0	57.1
	我们的方法（两阶段）	**100.0**	**100.0**	**100.0**
身份跟踪	无度量预训练	95.0	95.0	95.0
	联合端到端	**100.0**	**100.0**	**100.0**
	我们的方法（两阶段）	**100.0**	**100.0**	**100.0**
空间重构	无度量预训练	87.5	87.5	87.5
	联合端到端	81.0	85.0	82.9
	我们的方法（两阶段）	**90.0**	**90.0**	**90.0**

C. 视觉-语言-动作系统提示词详情

为了有效地对 K 关键帧链式连接视觉-语言-动作模型在时间上下文中，我们构建了一个结构化系统提示词，用于指导模型解读 交错图像序列。

由于我们的输入流包含历史关键帧和当前观测，且每个时间戳都包含多视角数据（第三人称视角和腕部视角），因此提示词必须明确规定这种结构，以避免歧义。我们在实验中使用的完整系统提示词见方框 5。

系统提示模板

任务：**{语言指令}**
注意：所提供的图像是一系列按时间顺序排列的**图像对**。每个时间步提供两张图像：第一张是**第三人称视角**，第二张是对应的**腕部视角**。最后一组图像代表当前状态，而此前所有图像对均为历史记录。根据这一多视角序列，分析机器人的进展，并确定下一步的即时动作。

图 5.所使用的结构化系统 提示，用于条件化视觉-语言-动作策略。{语言指令} 占位符会被替换 d
为具体的自然语言 本回合的指令（例如："根据立方体最初的垂直顺序堆叠立方体"）。

D. 定性真实世界可视化

为了具体展示系统在物理环境中的行为，我们在图中呈现了全部四种非马尔可夫任务的逐帧执行序列6。

这些可视化展示了智能体如何应对各类依赖记忆的挑战：

- **空间重构：**堆叠被打乱后，智能体成功恢复了初始方块排列。
- **时序排序：**策略严格遵循指令所限定的 $Red \rightarrow Green \rightarrow Blue$ 排序顺序。
- **计数与延迟：**机器人在信号间隔期间正确暂停，并仅在第二次闪光之后执行推。
- **身份跟踪：**尽管方块外观相同，智能体仍能在人为操作的交换之后正确识别并取回目标方块（最初位于中心）。

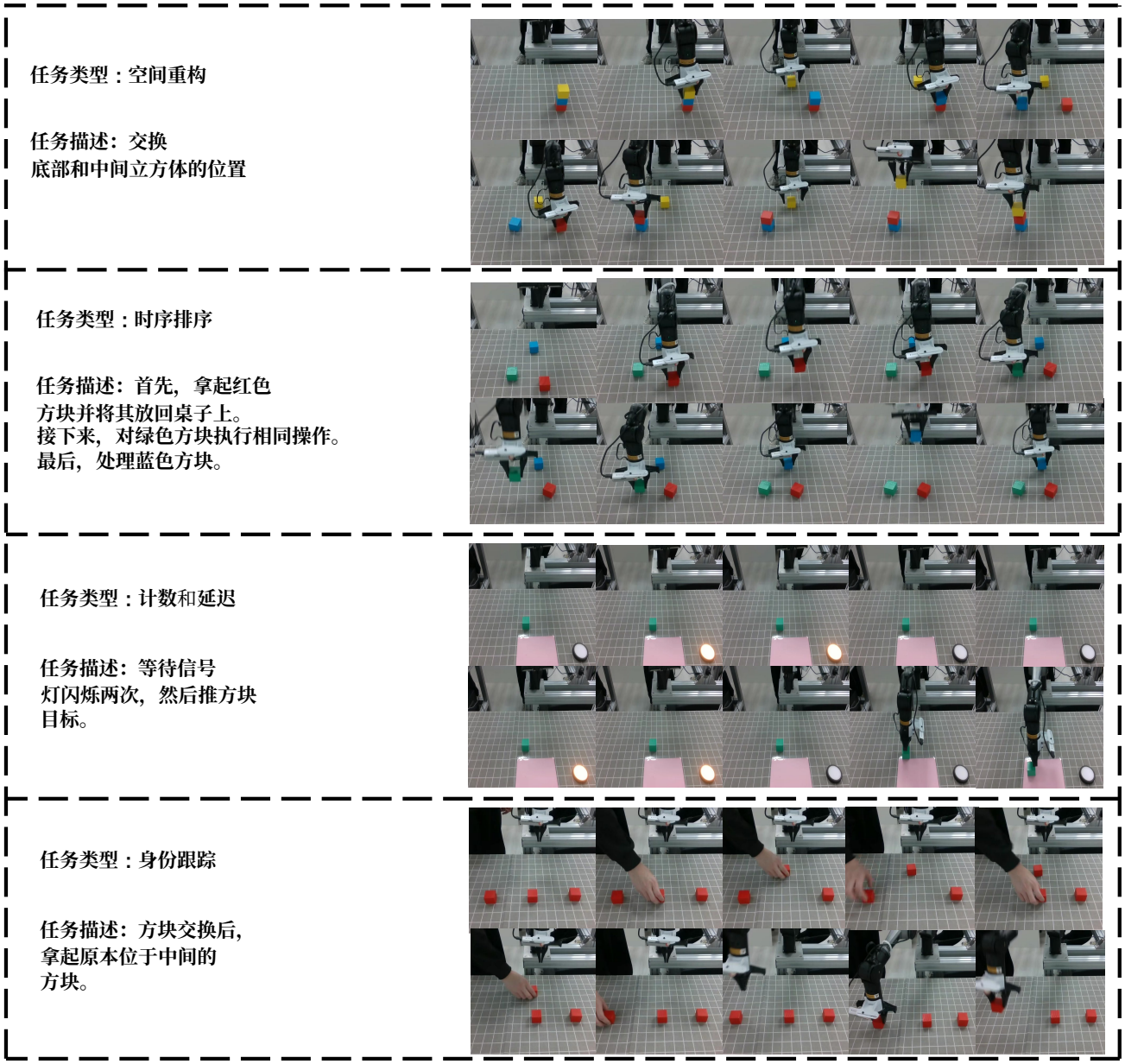

图 *6.* **关键帧链式视觉-语言-动作模型在真实场景中的定性执行过程。** 每一行对应四个非马尔可夫任务之一。快照展示了智能体维持长时程一致性的能力，例如等待信号灯（第 3 行）或跟踪交换后的物体（第 4 行）。

E. 实现细节和超参数

表 *6.* **VLA微调**

参数	值
基础模型	GR00T-N1.5-3B
优化器	AdamW
LR	1×10^{-4}
权重衰减	1×10^{-5}
批量大小	16（每个 GPU）
梯度累积	1
最大步数	50,000
预热	0.05
调度器	余弦
后端	TorchCodec
可训练	
投影器	✓
扩散头	✓
骨干网络	冻结

表 *7.* **KSM 阶段 I**

参数	值
批量大小	64
训练轮数	30
优化器	AdamW
LR	1×10^{-4}
权重衰减	1×10^{-3}
Loss	三元组间隔
间隔 (m)	1.0
度量	欧几里得（L_2）

表 *8.* **KSM第二阶段**

参数	值
上下文（k）	3帧
批量大小	32
训练轮数	50
优化器	AdamW
LR	1×10^{-4}
权重衰减	0.05
Loss	带 logits 的二元交叉熵
正类权重	5.0