arXiv:2602.01018v1 [cs.RO] 1 Feb 2026

从多任务轨迹中离线发现可解释技能

Chongyu Zhu[1†*]、Mithun Vanniasinghe[2†]、Jiayu Chen[3],和 Chi-Guhn Lee[1]

摘要—分层模仿学习是一种强大的范式，可用于从演示中习得复杂的机器人行为。然而，一个核心挑战在于如何从长时域、多任务离线数据中发现可复用技能，尤其是在数据缺少显式奖励或子任务标注的情况下。本文提出 LOKI，这是一个端到端的三阶段学习框架，旨在实现离线技能发现和分层模仿。该框架首先进行两阶段的弱监督技能发现：第一阶段利用弱任务标签引导的、带对齐约束的向量量化变分自编码器，执行粗粒度、任务感知的宏观分段；第二阶段则使用自监督序列模型在微观层面细化这些片段，随后通过迭代聚类来整合技能边界。第三阶段利用这些精确的边界，在基于选项的框架中构建分层策略，并学习终止条件 β 以实现显式技能切换。LOKI 在具有挑战性的 D4RL Kitchen 基准上取得了较高的成功率，性能优于标准分层模仿学习基线。此外，我们还证明，发现的技能在语义上有意义，符合人类直觉，并且具有组合性：通过成功地将这些技能串联起来，系统能够解决一个新颖且未见过的任务。代码可在此处获取：
https://github.com/cyzhu-hiter/LOKI

I. 引言

让智能机器人系统能够在多样且非结构化的环境中执行复杂的长时域任务，是现代人工智能和机器人学研究的核心目标之一 [1] 。无论是依赖精细的人工编程和基于规则的控制，还是在强化学习中设计具有挑战性的稠密奖励函数，传统方法在面对这类复杂任务时，往往在可扩展性、适应性和可调试性方面遭遇重大障碍 [2], [3]。对此，模仿学习（IL）应运而生，成为一种极具前景的范式，使机器人能够直接从人类演示中习得行为。模仿学习有效避开了显式策略工程和稀疏奖励信号带来的复杂性，因此对于无法直接编程或编程成本高昂的任务尤为重要 [3], [4]。在此基础上，分层模仿学习（HIL）进一步推动了该领域的发展，它将复杂任务分解为易于管理的子任务，从而简化控制并促进长时域行为的学习。分层模仿学习已在多个领域取得显著成效，从复杂的机器人操作和装配到稳健的自主导航；与全面的通用模型相比，它也日益受到青睐，因为它能够在资源受限的计算单元上部署模块化策略 [5], [6], [7]。

然而，将模仿学习扩展到真正复杂的多任务场景仍面临重大挑战。现实世界中的人类示范虽然信息丰富，但通常是在离线条件下收集的，缺少明确的子任务标签，而且包含长而未分段的动作序列，涵盖多种不同的行为。如何从此类原始、未经标注的离线多任务数据中发现一组可复用且语义有意义的技能，仍然是一项艰巨的挑战 [3]。现有方法往往难以解决若干关键问题：（1）技能边界模糊：在没有明确监督的情况下，难以将连续轨迹准确划分为离散且具有明确目的的技能，导致技能定义不够理想 [8]。（2）缺乏可复用性与可解释性：学习得到的技能可能高度依赖特定任务，或无法契合人类的概念理解，从而妨碍其迁移到新任务或用于人机协作 [3], [9]。（3）难以扩展至多任务设置：许多方法是为单任务设置设计的，或需要大量特定任务工程，因此难以应用于多样化的多任务环境，而这类环境迫切需要具备泛化能力的技能 [10]。尽管分层模仿学习（HIL）提供了一种颇具前景的框架，可通过组合更简单的技能来处理长时域任务，但其有效性从根本上取决于所发现的底层技能的质量与语义连贯性 [11]。

本文提出了通过模仿学习离线关键技能（LOKI）方法，以弥补这些关键不足。LOKI 是一种新颖的三阶段端到端学习框架，可从长时程、多任务离线数据中实现稳健的技能发现与分层模仿。该设计受到两阶段技能发现理念的启发，这一理念既有效，也符合人类直觉，如图 1 所示。受向量量化变分自编码器（VQ-VAE） [12], 的启发，我们的框架首先利用一种新颖的对齐约束 VQ-VAE（EVQ-VAE）实现通用技能分解。随后，通过任务技能对齐聚类过程进一步细化所得的分段结果。该方法使机器人能够直接从原始演示中学习稳健、可解释且可组合的技能集，为构建能力更强的智能体机器人系统铺平道路。我们的主要贡献如下：

- 我们提出了 LOKI，这是一种新颖的三阶段弱监督框架，可从长时域、多任务离线数据中发现技能，并支持构建可执行的、基于选项的分层策略。

†同等贡献。[1]C. Zhu 和 C.-G. Lee 任职于加拿大安大略省多伦多市的多伦多大学机械与工业工程系，以及运筹学与强化学习（DORL）实验室。[2]M. Vanniasinghe 任职于加拿大安大略省多伦多市的多伦多大学航空航天研究所（UTIAS）。[3]J. Chen 任职于中国香港特别行政区的香港大学智能体智能实验室。*通讯作者：
cyzhu@mie.utoronto.ca

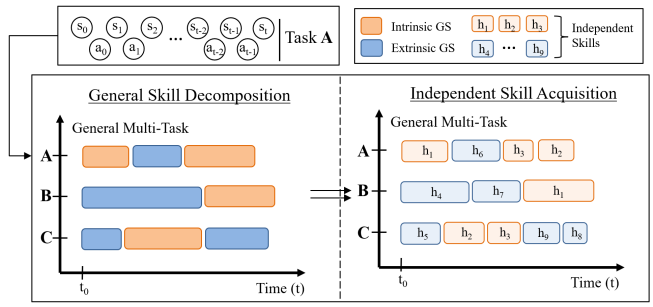


图1. 两阶段技能发现：从长时域多任务轨迹到可复用独立技能。传统的纯技能发现通过先验知识 [6] 或信息论 [13] 处理该问题时往往难以奏效，而我们的方法首先将一条轨迹处理为任务相关的外在通用技能（GS）和与任务无关的内在通用技能。随后，通过弱监督对齐，将这些技能分解为细粒度、与人类知识对齐的独立技能。该方法能够实现稳健的技能发现，并便于针对新任务灵活重组技能。

- 我们提出了 EVQ-VAE，这是一种新颖的架构，利用弱任务标签进行粗粒度的任务感知宏观分段。
- 我们提出一种弱监督微观分段方法，以区分保守技能。
- 我们证明了该框架的有效性，所发现的技能具有高度可解释性与一致性，并且可组合以解决新任务。

II. 相关工作

A. 技能发现

技能发现旨在从非结构化轨迹数据中提取一组可复用且有意义的行为原语，这是在机器人学中实现分层控制和长时域任务规划的基石。该领域可大致分为在线方法与离线方法。在线方法通常通过与环境的直接交互来学习技能，往往通过最大化状态与潜在技能变量之间的互信息，以鼓励多样且可预测的行为 [13]。其他在线方法则侧重于学习能够到达状态空间不同区域的技能，强调可达性与探索 [14], [15]。尽管这些方法功能强大，但需要大量且往往代价高昂的在线交互，这在真实世界的机器人学中并不总是可行。

相比之下，离线技能发现基于固定的预采集轨迹数据集开展。这种设定更具实用性，但也更具挑战性，因为智能体无法通过反复试错来纠正自身行为，而是完全依赖所提供数据的质量和覆盖范围 [16]。在方法论上，离线方法通常类似于无监督聚类。一种重要的研究思路是利用离散生成模型，例如向量量化变分自编码器（VQ-VAE），学习一个与不同技能隐式对应的离散潜在空间 [12], [17]。另一种方法则将技能发现视为轨迹分段问题，把轨迹作为时间序列数据，并试图识别出表明行为之间发生状态转移的变点 [18]。

B. 分层模仿学习

分层模仿学习是一种解决长时域任务的有效范式，它将复杂的控制问题分解为两层结构：以时间抽象的方式做出决策的高层策略，以及将这些决策执行为具体动作的低层策略 [3]。它为利用通过技能发现识别出的可复用技能提供了自然的框架。低层策略通常是一个以技能为条件的控制器，它将当前状态和来自高层策略的潜在代码作为输入，以生成动作。它通常通过行为克隆 [19] 进行训练，并在达到固定时间步数或收到终止信号时结束 [20]。

分层模仿学习研究的重点通常在于高层策略的设计与学习。一种常见方法是将神经网络训练为动态元控制器，使其学习根据当前状态选择最优技能 [3]。尽管这种方法灵活，但可能较为复杂且样本效率低。另一种方法是采用结构更明确的高层策略，例如任务图或固定技能序列，这些策略可以根据先验知识推导，也可以直接从示范数据中提取 [3], [8], [21]。

C. 视觉-语言-动作模型

视觉-语言-动作（VLA）模型近来已成为机器人学中的主流范式，这在很大程度上得益于大语言模型（大语言模型）的进展 [1], [5]。这些模型利用基础模型强大的泛化能力，将高层输入（如自然语言指令或视觉场景）直接映射为动作序列 [22], [23]。然而，VLA的成功在很大程度上依赖于能否获取带有丰富语言与文本标注的海量网络规模数据集，即便是在预训练基座上进行微调时亦然。

在控制层级的语境中，VLA通常作为高层元控制器运行。它们擅长语义推理与决策——决定下一步要执行何种技能（例如，“拿起红色方块”），而不是指定执行该技能所需的细粒度关节级运动 [5]。我们的工作与这一范式互补，侧重于离散技能的低层原语学习与执行，而VLA则侧重于将语言映射为行为序列。

III. 方法

本节将详细介绍 LOKI——一种用于离线技能发现和分层模仿的新型三阶段端到端学习框架。我们将详细说明其架构、学习目标，以及针对现有技术所作的独特调整。LOKI 的完整工作流程如图 2 所示，其中直观展示了原始轨迹如何被转化为经过细化的技能，并最终驱动可执行的分层策略。

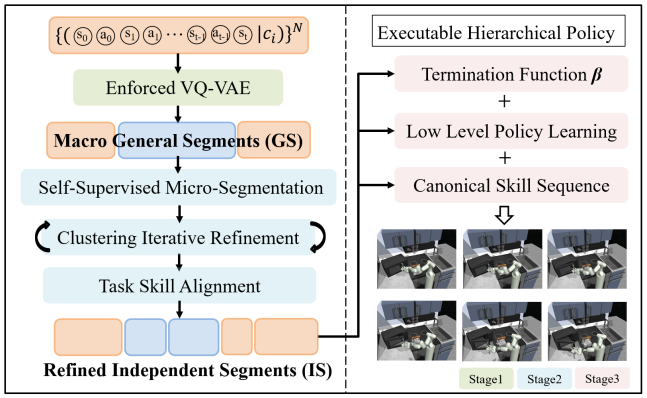


图 2. LOKI 框架示意图。左图展示了两阶段技能发现与分割过程：（浅绿色）使用强制VQ-VAE进行宏观分段，随后是（浅蓝色）结合聚类迭代细化的自监督微观分割，以及任务技能对齐，从而得到细化的独立片段（IS）。右图描绘了该框架的第三阶段：根据所发现的技能构建可执行的分层策略，其中包括学习终止函数（β）、低层策略，并推导用于任务执行的规范技能序列策略。技能的平滑执行以可视化输出形式呈现。

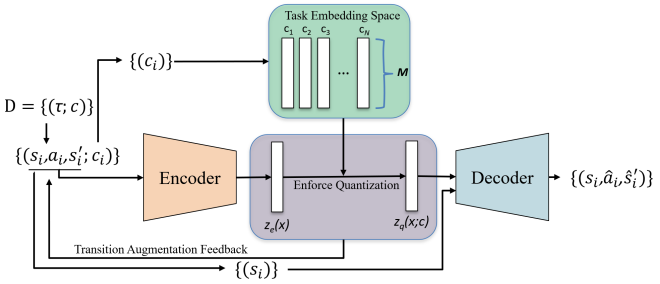


图 3. EVQ-VAE 架构示意图，该架构利用弱任务标签 c_i，引导嵌入 $z_e(x)$ 到码本向量 $z_q(x;c)$ 的强制量化。该机制能够自然地对任务相关的状态和动作进行聚类，同时为多任务行为保留多样化的表示，从而实现细粒度的轨迹分解。M 是码本向量的维度。

A. 阶段 *1*：通过强制*VQ-VAE*进行宏观分段

设我们的离线数据集为 $\mathcal{D} = \{(\tau_i, c_i)\}_{i=1}^{N}$，其中每条轨迹 τ_i 由一系列状态转移组成，且 $c_i \in \{1, \ldots, M\}$ 是一个弱标签，用于指示该轨迹所属的高层任务。每个状态转移定义为一个元组 $x_t = (s_t, a_t, s_{t+1})$。

标准的 VQ-VAE [12], [24] 由编码器 $\mathcal{E}$、解码器 $\mathcal{D}$ 以及可学习的码本 $E = \{e_k\}_{k=1}^{K}$ 组成。编码器将输入 x_t 映射为连续的潜在嵌入$z_e(x_t) = \mathcal{E}(x_t)$。随后，该嵌入通过替换为码本中最近的向量而被量化，$z_q(x_t) = e_{k^*}$，其中 $k^* = \text{argmin}_k \|z_e(x_t) - e_k\|_2$。VQ-VAE 的总损失函数如下：

$$\mathcal{L}_{\text{VQ-VAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{codebook}} + \lambda \mathcal{L}_{\text{commit}} \tag{1}$$

这里，重构损失 $\mathcal{L}_{\text{recon}}$ 确保输入得到准确重构。码本损失 $\mathcal{L}_{\text{codebook}}$ 会更新码本嵌入，使其与编码器输出相匹配；而承诺损失 $\mathcal{L}_{\text{commit}}$ 则促使编码器输出归属于特定的码本向量，从而防止码本波动；λ 是承诺项的加权超参数。

1) 基于潜在空间熵的边界检测：训练好的 EVQ-VAE 提供了一种有依据的方法来检测技能边界。对于给定的转移 x_t，我们计算其连续嵌入 $z_e(x_t)$ 与每个码本向量 e_k 之间的 L2 距离。随后，通过 softmax 函数对这些距离进行归一化，得到码本上的概率分布：

$$p_t(k) = \frac{\exp(-\|z_e(x_t) - e_k\|_2^2)}{\sum_{j=1}^{K} \exp(-\|z_e(x_t) - e_j\|_2^2)} \tag{2}$$

随后，我们计算该分布的熵，$H(p_t) = -\sum_{k=1}^{K} p_t(k) \log p_t(k)$。

这种熵是用于分段的强大信号。外在技能是特定任务独有的，因此会产生低熵分布，因为 $z_e(x_t)$ 非常接近与任务对齐的码本向量。相反，内在技能在多个任务中普遍存在，其嵌入与多个码本向量的距离大致相等，因此会产生高熵分布。沿轨迹出现的熵信号显著变化表明存在变点，这些变点标示了内在通用技能与外在通用技能之间的边界。我们使用 ruptures 库[25]中的 RBF 核 PELT 算法，稳健地识别这些变点。

2) 强制*VQ-VAE*（*EVQ-VAE*）：尽管标准VQ-VAE在离散化方面有效，但它缺乏将潜在编码与特定任务结构对齐的固有机制，而这对于多任务设置中的可解释技能发现至关重要。为克服这一局限并实现任务条件技能学习，我们引入强制矢量量化变分自编码器（EVQ-VAE），如图3所示。我们的EVQ-VAE引入了两项关键改进：（1）利用弱任务标签 c来强制执行任务条件量化；（2）引入一种新颖的散度损失，以促进解耦码本表示。

给定输入转移 $x_t = (s_t, a_t, s_{t+1})$及其关联的任务标签 c，EVQ-VAE旨在重构动作和下一状态$(\hat{a}_t, \hat{s}_{t+1})$。EVQ-VAE的总损失函数定义为：

$$\begin{aligned}\mathcal{L}_{\text{EVQ-VAE}} =& \mathcal{L}_{\text{recon}}(x;c) + \mathcal{L}_{\text{codebook}}(x;c) \\ &+ \beta \mathcal{L}_{\text{commit}}(x;c) + \gamma \mathcal{L}_{\text{divergence}}\end{aligned} \tag{3}$$

损失项详述如下。

任务条件重构损失($\mathcal{L}_{\text{recon}}(c)$)旨在促使解码器根据任务条件量化潜在表示 $z_q(x_t;c)$和任务标签 c，准确重构输入(a_t, s_{t+1})：

$$\mathcal{L}_{\text{recon}}(x_t;c) = \|a_t - \hat{a}_t\|^2 + \|s_{t+1} - \hat{s}_{t+1}\|^2 \tag{4}$$

任务条件化的码本损失 $\mathcal{L}_{\text{codebook}}(x_t;c)$ 和承诺损失 $\mathcal{L}_{\text{commit}}(x_t;c)$ 经过调整，以确保码本向量和编码器输出在任务条件化下保持一致地学习

。关键在于，量化现在会显式考虑任务标签 c。与简单地选择全局最近的码本向量不同，约束机制确保所选的 $z_q(x_t; c)$ 是与任务 c 相关的最近码本向量。这种引导会隐式地将编码器输出 $z_e(x_t)$ 推向特定于任务的嵌入，并相应地更新码本。具体形式如下：

$$\mathcal{L}_{\text{codebook}}(x_t; c) = \|\text{sg}[z_e(x_t)] - z_q(x_t; c)\|_2^2 \tag{5}$$

$$\mathcal{L}_{\text{commit}}(x_t; c) = \|z_e(x_t) - \text{sg}[z_q(x_t; c)]\|_2^2 \tag{6}$$

其中 sg[·] 是停止梯度算子

引入码本散度损失（$\mathcal{L}_{\text{divergence}}$），以鼓励码本向量之间形成彼此不同且专门化的表示，从而促进发现独特的外在技能并防止码本坍塌。它被定义为所有唯一码本向量对之间欧几里得距离平方和的负值；γ 是该项的加权系数：

$$\mathcal{L}_{\text{divergence}} = -\sum_{k=1}^{K}\sum_{l=k+1}^{K} \|e_k - e_l\|_2^2 \tag{7}$$

B. 第二阶段：弱监督微观分段

在使用 EVQ-VAE 将轨迹宏观分段为通用技能（GS）之后，我们的第二阶段着重于将这些宏片段进一步分解为更细粒度、与上下文无关的独立技能（IS）。这种微分割通过弱监督方法实现，以高精度细化技能边界，并建立在 OPAL 框架 [17]的原则之上。

为此，我们采用序列变分自编码器架构。我们并不处理长度可变的整个宏片段，而是使用固定大小的滑动窗口来创建训练样本。给定一个宏片段，我们生成一组固定长度的窗口 $\mathcal{D}_{\text{win}} = \{\tau_{\text{win}}\}$，其中每个窗口 $\tau_{\text{win}} = \{(s_t, a_t, s_{t+1})\}_{t=i}^{i+W-1}$ 的长度为 W。该模型的目标是精确重建这些窗口，并以与先验分布一致的方式组织其潜在空间，该先验分布以窗口的初始状态和弱任务标签为条件。总损失定义为：

$$\mathcal{L}_{\text{micro}} = \mathbb{E}_{\tau_{\text{win}}\sim\mathcal{D}_{\text{win}}}\left[\mathcal{L}_{\text{recon}} + \alpha_{\text{KL}}\mathcal{L}_{\text{KL}}\right] \tag{8}$$

重构损失 ($\mathcal{L}_{\text{recon}}$) 促使模型通过重构窗口内的动作 a_t 和后续状态 s_{t+1}，捕捉细粒度的动力学特征：

$$\mathcal{L}_{\text{recon}} = \sum_{t=i}^{i+W-1} \left(\|s_{t+1} - \hat{s}_{t+1}\|_2^2 + \|a_t - \hat{a}_t\|_2^2\right) \tag{9}$$

Kullback-Leibler（KL）散度项（$\mathcal{L}_{\text{KL}}$）对从整个窗口编码得到的后验分布 $q(z|\tau_{\text{win}})$进行正则化，使其接近先验分布 $p(z|s_i, c)$。该先验仅以窗口的初始状态s_i 和任务标签 c为条件。这一设计对于分层执行至关重要，因为在分层执行中，策略无法观察未来。它确保代表一项技能并概括未来行为序列的潜在代码 z，能够被仅依赖于当前状态的先验合理地近似。KL散度表示为：

$$\mathcal{L}_{\text{KL}} = \text{KL}(q(z|\tau_{\text{win}})||p(z|s_i, c)) \tag{10}$$

1) 技能稳定性与边界检测：不同技能的一个关键特征是其时间稳定性。当智能体执行明确定义的技能时，我们的模型能够高精度地重构其状态和动作转移，因此重构误差较低。相反，偏离当前技能预期的动力学通常意味着可能发生技能变化或出现边界。因此，我们发现，重构误差信号的突然峰值能够可靠地指示潜在的技能转换。为了稳健地识别这些峰值，我们分析误差曲线的局部几何特征。具体而言，我们将平滑后的误差信号的二阶导数出现显著局部峰值所对应的时间步确定为候选分割点。这些峰值准确捕捉向上凹性最大的点，对应于“尖峰”的起始位置，从而使我们能够在宏片段内生成一组初始的潜在技能边界。

2) 微分割的迭代细化：由误差信号生成的初始候选分割点可能仍然存在噪声或过于细碎。为了获得功能上彼此 distinct 的独立技能，我们采用迭代细化过程，系统地合并冗余片段。该过程首先根据当前的候选分割点集合，将一条轨迹划分为更小的片段。随后，使用训练好的序列变分自编码器编码器，将每个片段编码为潜在嵌入。接着，对所有片段嵌入应用 K-Means 聚类算法，并为每个片段分配一个簇标签。当同一轨迹中的两个相邻片段被分配到同一个簇时，便执行关键的细化步骤：将分隔二者的候选分割点视为冗余点并移除。该迭代合并过程仅应用于微分割阶段识别出的候选分割点，而宏观分段阶段确定的“硬”边界则予以保留，以维护整体任务结构。该过程会重复固定轮数，或持续进行直至分割点集合收敛。

3) 任务技能对齐：在获得一组稳定的优化后切分点后，技能发现的最后一步是为每个片段分配有意义且一致的技能 ID。我们采用任务技能对齐策略，区分可复用、与任务无关的（内在）技能和独特、特定于任务的（外在）技能。首先，我们仅对所有已识别内在片段的潜在嵌入执行 K 均值聚类，将它们归为一组可复用技能基元 $\{k_1, \ldots, k_{K_{\text{int}}}\}$。相比之下，每种不同的外在技能都会从另一组 $\{k_{K_{\text{int}}+1}, \ldots, k_K\}$ 中直接分配一个唯一且不重叠的技能 ID，并将其视为无法跨任务迁移的专用基元。最后，为确保结构一致性，我们要求属于同一

任务的所有轨迹必须共享完全相同的技能序列。对于每项任务，我们通过分析其演示中观测到的序列之间的相似度和出现频率，确定最可能的或“典型”技能序列。随后，将该任务的所有轨迹强制对齐到这一典型序列。由此得到最终的结构化数据集，其中每个片段都标记有语义连贯且一致的技能 ID，可用于策略学习的最后阶段。

C. 分层策略学习

一旦将轨迹分割为一组离散的、相互独立的技能，我们框架的最终阶段便是构建一个能够执行这些技能以解决指定任务的分层策略。该策略由三个关键组件构成：学习得到的基于状态的终止函数、技能条件低层策略，以及确定性的高层任务程序。

1) 终止函数学习：为了管理技能之间的转换，我们学习了一个基于状态的终止函数，$\beta(s) \rightarrow [0,1]$，该函数改编自选项框架[20]。不同于依赖固定技能时间范围的方法，我们的方法根据是否到达具有意义的状态边界来学习终止技能。我们将其表述为一个二分类问题：将每个已发现技能片段的最终状态标记为正例（表示终止），并将技能片段中的所有其他状态标记为负例。在这些数据上训练一个多层感知机，以学习函数 $\beta(s)$。在执行过程中，当 $\beta(s_t)$ 的输出超过预定义阈值时，技能将终止。

2) 通过条件行为克隆训练低层策略：每个原子技能的执行均由一个统一的低层策略负责，$\pi(a_t|s_t,\ k)$，该策略通过条件行为克隆进行训练。策略以当前状态 s_t 和目标技能 ID 的独热编码 $k \in \{1,\ldots,K\}$为条件。训练数据集由经过优化的技能片段中的所有状态-动作对组成，每对 $(s_t,\ a_t)$均标记有相应的技能 ID k。随后，使用标准监督学习目标训练策略，以最大化演示动作的对数似然：

$$\mathcal{L}_{\text{BC}} = \sum_{(s_t,a_t,k)\in\mathcal{D}_{\text{segments}}} -\log \pi(a_t|s_t,k) \qquad (11)$$

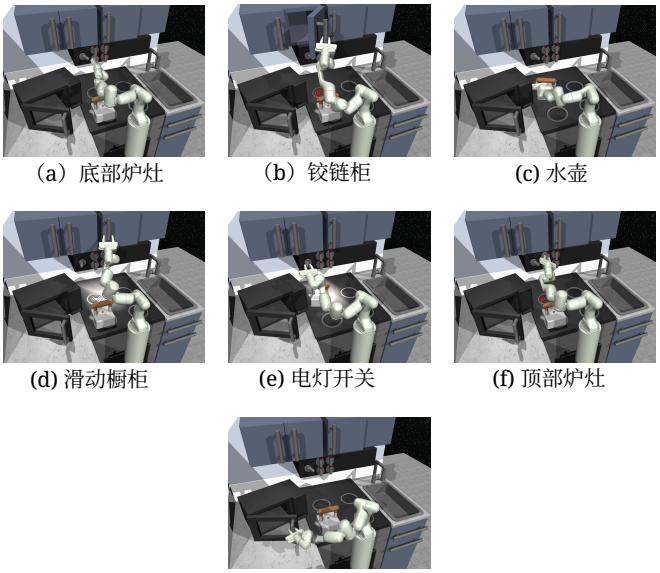
其中，$\mathcal{D}_{\text{segments}}$ 是第二阶段发现的所有技能片段的完整数据集，详见第 III-B 节。

3) 典型技能序列作为高层策略：我们框架中的高层策略并非通过学习得到的动态控制器。相反，我们利用第二阶段发现并对齐的高度一致的技能序列（见第 III-B.3 节），将其作为确定性的高层任务程序。在执行过程中，智能体只需遵循针对给定任务预先定义的序列，并在当前技能完成后转至下一个技能。这种方法大幅简化了学习问题，将技能选择的需求转化为直接执行预先确定的脚本，从而使控制结构更加稳健且易于理解。

IV. 实验

A. 实验设置

1) 环境与数据集：我们的实验在具有挑战性的 Franka Kitchen 操作环境中进行，这是长时程、多任务机器人学习的标准基准。我们使用 D4RL 基准中的“kitchen-mixed-v0”离线数据集 [26], ，该数据集由接力策略学习框架 [4]生成。这个全面的数据集包含丰富的行为，涵盖七个不同的子任务，包括与微波炉、水壶、电灯开关以及各种橱柜等物体交互。这些原子子任务在图 4 中以可视化方式呈现。为清晰起见，我们在后续讨论中用一个代表性词语指代每个子任务。



（a）底部炉灶　（b）铰链柜　(c) 水壶
(d) 滑动橱柜　(e) 电灯开关　(f) 顶部炉灶
(g) 微波炉

图 4. D4RLKitchen 数据集 [26] 中不同子任务的概览。完整的轨迹数据集包含七个子任务：(a) 打开底部炉灶；(b) 打开铰链柜；(c) 移动水壶；(d) 打开滑动橱柜；(e) 打开电灯开关；(f) 打开顶部炉灶；(g) 打开微波炉，并且每条轨迹包含四个子任务。

2) 任务选择与数据结构：从完整的 D4RL 数据集中，我们为三个特定的复杂多阶段任务各选取约 24 条成功的演示轨迹，以评估我们的方法。任务如下，我们用首字母缩写来简要表示这些任务：

- 微波炉 水壶 开关 铰链 - MKSH
- 微波炉 底部 顶部 铰链 - MBTH
- 微波炉 开关 滑动 铰链 - MSSH

这个经过筛选的数据集共包含 71 条成功轨迹，构成了我们实验评估的基础。在七个子任务中，微波炉、铰链和开关是内在技能，而水壶、底部、顶部和滑动是外在技能。

3) 数据特征与监督：我们的离线数据集因其固有特征而面临重大挑战。每条轨迹仅由一系列状态-动作对 ($s_t,\ a_t$) 组成，且尤为重要的是，不包含奖励信号，也没有关于子任务边界的显式标注。

这些演示采集自成功执行的过程，这意味着每条轨迹都实现了其最终任务目标。我们提供的唯一监督形式是为每条轨迹提供一个弱任务标签，指出其所属的三个复杂任务之一。这种最小监督设置旨在严格检验我们的模型在无需大量人工标注或稠密奖励信号的情况下，发现有意义且可复用技能的能力。

B. 结果与分析

1) 宏分割分析：图 5 展示了我们的强制VQ-VAE（EVQ-VAE）在三个复杂任务的代表性轨迹上得到的宏观分段结果。结果表明，我们的弱监督方法成功识别了不同通用技能之间的边界。图中显示了轨迹进程中码本概率分布的对数熵。正如第 III-A.1 节所假设的，熵信号是区分内在技能与外在技能的可靠指标。

我们在所有任务中都观察到一致的规律。在外在技能期间（这类技能是特定任务独有的），转移嵌入与单个任务专属码本向量高度对齐，因此呈现低熵状态。相反，在多个任务共有的内在技能期间，嵌入与多个码本向量的距离更加接近，因此呈现高熵状态。例如，在 MKSH 任务中，EVQ-VAE 正确地将水壶子任务识别为低熵的外在技能，而周围的子任务则呈现出内在技能的高熵特征。这种清晰的区分验证了 EVQ-VAE 能够稳健地进行宏观分割，有效地将复杂、未标注的轨迹分解为基础的通用技能序列，以便后续进行微观层级细化。

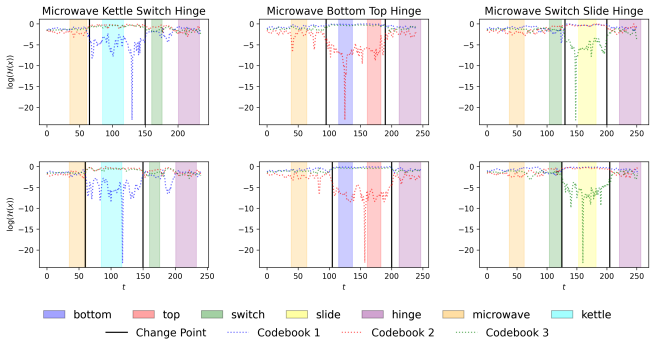


图 5. 我们的 EVQ-VAE 在三个复杂任务上的变点检测结果。纵轴表示对数熵，该值根据某个转移的嵌入与各码本向量之间的 L2 距离计算得出。不同颜色（蓝色、红色、绿色）的虚线分别对应嵌入属于三个主要码本向量之一的对数概率。低熵状态下，某个码本向量占主导地位（一条虚线明显下降），表示外在（特定于任务的）技能。高熵状态下，各线相互交错，没有单一码本占主导地位，表示内在（与任务无关的）技能。彩色阴影区域表示各子任务的真值持续时间，黑色实线竖线则表示我们的方法检测到的变点。

2）微分割分析：为评估第二阶段微分割过程的有效性，我们分析迭代细化前后所发现技能边界的质量。图 6 展示了三个评估任务中具有代表性的宏观片段之间的对比。根据重构误差信号生成的初始分割点（绿色虚线）成功识别了大多数主要技能转换。然而，这些分割点可能带有噪声或过于细碎，常常在单个连续子任务内部设置多个冗余边界（例如“任务 MKSH Seg0 原始”）。

我们的迭代细化和任务对齐机制有效地解决了这些问题。如“细化后”行所示，该过程通过合并被分配到同一技能簇的连续片段，成功移除了无效的分割点。例如，在“Task MKSH Seg0 Refined”图中，“微波炉”子任务内部最初的多个分割点被正确合并。此外，任务对齐组件确保属于同一高层任务的所有轨迹都收敛到一致的分段结构。此过程不仅能移除不必要的边界，还能在某条轨迹遗漏了必要的分割点、而同一任务类型的其他轨迹中存在该分割点时，将其补上。最终细化后的分割点（紫色虚线）呈现出清晰、稳健且一致的分段效果，与真实标注的子任务时长高度吻合。

3）技能与人类知识对齐：为验证所发现技能的质量和可解释性，我们对所有经过细化的独立技能片段的潜在空间进行可视化。图7展示了这些片段嵌入的二维主成分分析投影。可视化结果表明，尽管对应于同一真实标注子任务的片段嵌入存在相当大的差异——这反映了人类示范中自然存在的多样性和噪声——我们的聚类过程仍成功为每项核心技能识别出了稳定且彼此分离良好的质心（黑边星形标记）。

至关重要的是，学到的潜在空间展现出有意义的语义结构，并与人类知识相契合。涉及相似机器人运动或物体交互的技能，在空间中的位置彼此更近。例如，「微波炉」和「铰链柜」的簇都涉及手臂旋转运动，因此在空间中占据相邻区域。这种语义上的接近对分层控制大有裨益，因为它能够促进功能相关技能之间更平滑、更稳健的状态转移。这一结果证实，我们的框架不仅能有效地分割轨迹，还能学到结构清晰、可解释且符合人类任务概念的技能表征。

4) 终止函数：我们分层策略的有效性取决于终止函数 $\beta(s)$能否准确发出技能完成的信号。为评估这一点，我们在状态的留出测试集上，将其作为二元分类器来衡量其性能。我们将正确预测定义为并非单个时间步，而是真实技能边界周围的一个合理区间（ $\pm$4个时间步），以考虑执行中的轻微随机性。我们学习到的

表 I 基线性能比较（低层策略均方误差 ↓）。数值越低，模仿保真度越高。

方法	Task			平均值
	MKSH	MBTH	MSSH	
LOKI-O（我们的方法）	7.01±23.6	7.23±11.9	6.47±9.34	**6.94±15.3**
LOKI-E（我们的方法）	7.10±7.05	8.40±12.1	7.43±7.70	**7.76±9.69**
DADS	13.5±20.2	12.6±29.3	12.2±18.4	12.9±23.5
OPAL	11.6±24.1	12.6±21.7	9.79±23.0	10.6±22.9
专家（扁平行为克隆）	1.91±3.78	2.32±6.29	1.91±2.58	2.09±4.78

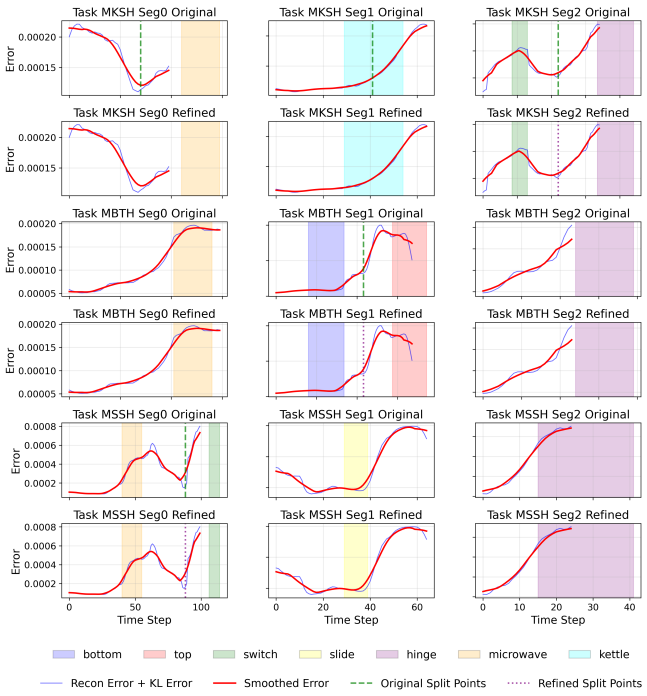


图 6. 微分段切分点的迭代细化过程结果。图中展示了三个评估任务的宏观分段的重构误差曲线。奇数行（“原始”）显示从误差信号中检测到的初始候选分割点（绿色虚线）。偶数行（“细化后”）显示经过迭代细化和任务对齐后的最终分割点（紫色点线）。每列代表相应任务中的一个不同宏观分段（Seg0、Seg1 等）。细化过程有效去除了含噪且冗余的分割点，同时确保同一任务的所有演示具有一致的分段结构。

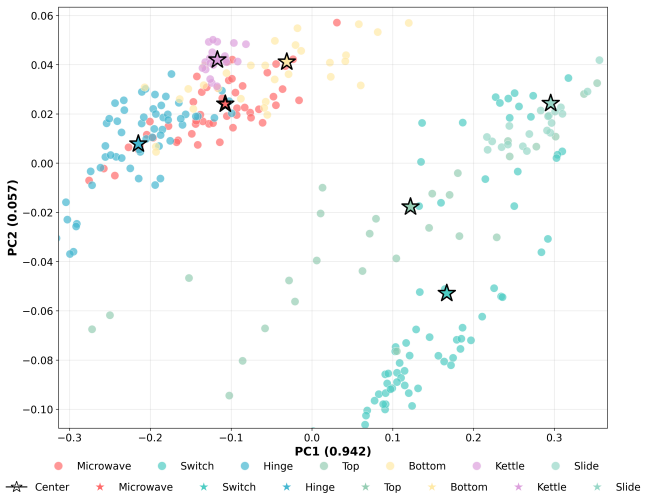


图 7. 发现的技能潜在空间的主成分分析可视化。每个点代表一个经过细化的独立技能片段，并根据其真实子任务标签着色。黑色描边的星形标记表示最终的 K 均值聚类中心，代表学到的典型技能。图中显示，尽管原始演示存在固有的噪声和方差（表现为相同颜色点的分散），我们的方法仍成功识别出了不同的聚类中心。某些聚类（例如微波炉和铰链）彼此接近，这表明该空间基于运动相似性学到了具有语义意义的结构。

终止函数在该既定区间内达到了93.7%的高准确率。这一结果表明，我们在分段阶段发现的技能边界对应于有意义且可学习的状态转移，从而为高层策略在不同技能之间切换提供稳健而可靠的信号。

5) 与基线方法的比较：我们使用留出的测试集上的一步超前预测均方误差（MSE），将我们的方法与若干基线方法进行比较。该离线指标有效评估了所学习的低层策略及其所依赖的底层技能表示的质量。结果汇总于表 I。

“专家”基线是一个扁平（非分层）的行为克隆（行为克隆，BC）模型，以高层任务标签为条件。其较低的均方误差（MSE）表明，该模型在分布内模仿方面表现出色，但缺乏基于技能的方法所具备的模块化和可解释性。我们提出的方法 LOKI 有两个变体：LOKI-O 使用独热技能 ID 作为条件，LOKI-E 则使用来自序列变分自编码器潜在空间的连续技能嵌入。

如表中所示，LOKI 的两个变体均显著优于分层基线方法 DADS 和 OPAL，这证明了我们的两阶段分段与任务对齐流程在发现更一致、更易模仿的技能集合方面具有优越性。值得注意的是，与 LOKI-O 相比，LOKI-E 的均方误差更低。我们将此归因于连续技能嵌入包含了更丰富的信息。独热 ID 将每项技能视为一个独立类别，而连续嵌入则能捕捉技能内部的细微差异以及不同技能之间的语义关系。这种更丰富的条件信号使低层策略能够生成更精准且更符合情境的动作。总体而言，这些结果验证了 LOKI

能够有效兼顾高保真模仿与结构化、可解释且可复用的技能词汇发现。

V. 结论与未来工作

本文解决了如何从缺少显式奖励或子任务标签的复杂多任务离线演示中发现可复用技能这一挑战。我们提出了 LOKI，这是一种三阶段框架，通过一种新颖的弱监督分割流程来应对这一问题。该流程首先识别粗粒度、任务感知的行为，随后将其细化为细粒度、相互独立的技能。在具有挑战性的 D4RL Kitchen环境中开展的实验表明，该方法效果显著。发现的技能不仅具有语义意义，并且与人类直觉高度一致，还使分层策略能够在任务复现中取得较高的成功率。总之，我们的工作提出了一种实用且鲁棒的方法，可从原始离线数据中提取结构化、可解释的技能，为实现更具可扩展性和泛化能力的从示范中进行机器人学习迈出了重要一步。

未来工作方面，我们确定了几个有前景的方向。首先，我们计划在更复杂、更多样化的机器人环境中应用并评估该框架，以进一步检验其可扩展性和鲁棒性。第二项关键扩展是调整框架，使其能够处理高维图像输入，而非结构化状态观测，从而显著拓宽其适用范围。最后，我们计划在其他标准离线技能发现基准上验证 LOKI 原理的泛化能力，进一步确立其作为从示范中进行机器人学习的通用工具的有效性。

参考文献

[1] P. Intelligence、K. Black、N. Brown、J. Darpinian、K. Dhabalia、D. Driess、A. Esmail、M. Equi、C. Finn、N. Fusai 等，“$\pi_{0.5}$：一种具备开放世界泛化能力的视觉-语言-动作模型”，*arXiv* 预印本 *arXiv:2504.16054*，2025。[2] H. Xue、C. Pan、Z. Yi、G. Qu 和 G. Shi，“通过扩散式退火实现力矩级运动控制的全阶基于采样的模型预测控制”，载于*2025 IEEE*机器人与自动化国际会议 *(ICRA)*。IEEE，2025，第 4974–4981 页。[3] M. Zare 、 P. M. Kebria 、 A. Khosravi 和 S. Nahavandi，“模仿学习综述：算法、最新进展与挑战”，*IEEE*控制论汇刊，2024。[4] A. Gupta、V. Kumar、C. Lynch、S. Levine 和 K. Hausman，“Relay策略学习：通过模仿学习和强化学习解决长时域任务”，*arXiv* 预印本 *arXiv:1910.11956*，2019。[5] Y. Zhong、F. Bai、S. Cai、X. Huang、Z. Chen、X. Zhang、Y. Wang、S. Guo、T. Guan、K. N. Lui 等，“视觉-语言-动作模型综述：从动作标记化的视角出发”，*arXiv* 预印本 *arXiv:2507.01925*，2025。[6] S. Venkatraman、S. Khaitan、R. T. Akella、J. Dolan、J. Schneider 和 G. Berseth，“离线强化学习中的潜在扩散推理”，*arXiv* 预印本 *arXiv:2309.06599*，2023。[7] S. Teng、L. Chen、Y. Ai、Y. Zhou、Z. Xuanyuan 和 X. Hu，“用于端到端自动驾驶的分层可解释模仿学习”，*IEEE*智能车辆汇刊，第 8 卷，第 1 期，第 673–683 页，2023。[8] X. Mao、G. Giudici、C. Coppola、K. Althoefer、I. Farkhatdinov、Z. Li 和 L. Jamone，“Dexskills：利用触觉数据进行技能分割，以学习自主长时域机器人操作任务”，载于*2024 IEEE/RSJ*智能机器人与系统国际会议 *(IROS)*。IEEE，2024，第 5104–5111 页。

[9] J. Wen、M. Zhu、Y. Zhu、Z. Tang、J. Li、Z. Zhou、C. Li、X. Liu、Y. Peng、C. Shen 等，“Diffusion-vla：通过自生成推理实现通用且可解释的机器人基础模型”，arXiv 预印本 arXiv:2412.03293，2024。[10] J. Chen、D. Tamboli、T. Lan 和 V. Aggarwal，“多任务分层对抗逆强化学习”，载于《机器学习国际会议论文集》。PMLR，2023，第 4895–4920 页。[11] R. W. Byrne 和 A. E. Russon，“模仿学习：一种分层方法”，《行为与脑科学》，第 21 卷，684，1998 5，667第。页–。[12] A. Van Den Oord、O. Vinyals 等，“神经离散表示学习”，《神经信息处理系统进展》，第 30 卷，2017。[13] B. Eysenbach、A. Gupta、J. Ibarz 和 S. Levine，“多样性即是所需：无奖励函数的技能学习”，arXiv 预印本 arXiv:1802.06070，2018。[14] A. Sharma、S. Gu、S. Levine、V. Kumar 和 K. Hausman，“动力学感知的无监督技能发现”，arXiv 预印本 arXiv:1907.01657，2019。[15] J. Chen 、 V. Aggarwal 和 T. Lan，“ 一种基于行列式点过程的无监督技能发现统一算法框架”，载于《神经信息处理系统进展》，2023。[16] S. Levine、A. Kumar、G. Tucker 和 J. Fu，“离线强化学习：教程、综述与开放问题展望”，arXiv 预印本 arXiv:2005.01643，2020。[17] A. Ajay、A. Kumar、P. Agrawal、S. Levine 和 O. Nachum，“OPAL：用于加速离线强化学习的离线基元发现”，arXiv 预印本 arXiv:2010.13611，2020。[18] S. Salimpour、J. P. Queralta 和 T. Westerlund，“自主巡检机器人的自校准异常与变化检测”，载于2022年第六届 IEEE机器人计算国际会议 (IRC)。IEEE，2022，第 207–214 页。[19] N. Esmaili、C. Sammut 和 G. Shirazi，“动态系统控制中的行为克隆”，载于1995 IEEE系统、人类与控制论国际会议：面向 21 世纪的智能系统，第 3 卷，1995，第 2904–2909 页，第 3 卷。[20] R. S. Sutton、D. Precup 和 S. Singh，“介于马尔可夫决策过程与半马尔可夫决策过程之间：强化学习中时间抽象的框架”，《人工智能》，第 112 卷，第 1 期，第 181–211 页，1999。[在线]。可访问：https://www.sciencedirect.com/science/article/pii/S0004370299000521[21] G. Vezzani、D. Tirumala、M. Wulfmeier、D. Rao、A. Abdolmaleki、B. Moran、T. Haarnoja、J. Humplik、R. Hafner、M. Neunert 等，“技能：面向高效时间延展探索的自适应技能排序”，arXiv 预印本 arXiv:2211.13743，2022。[22] B. Zitkovich、T. Yu、S. Xu、P. Xu、T. Xiao、F. Xia、J. Wu、P. Wohlhart、S. Welker、A. Wahid 等，“Rt-2：视觉-语言-动作模型将网络知识迁移至机器人控制”，载于《机器人学习会议论文集》。PMLR，2023，第 2165–2183 页。[23] S. Reed、K. Zolna、E. Parisotto、S. G. Colmenarejo、A. Novikov、G. Barth-Maron、M. Gimenez、Y. Sulsky、J. Kay、J. T. Springenberg 等，“通用智能体”，arXiv 预印本 arXiv:2205.06175，2022。[24] J. Chen、B. Ganguly、Y. Xu、Y. Mei、T. Lan 和 V. Aggarwal，“用于离线策略学习的深度生成模型：教程、综述与未来方向展望”，《机器学习研究汇刊》，第 2024 卷，2024。[25] K. Thurnhofer-Hemsi、E. L´opez-Rubio、M. A. Molina-Cabello 和 K. Najarian，“支持向量机分类器的径向基函数核优化”，arXiv 预印本 arXiv:2007.08233，2020。[26] J. Fu、A. Kumar、O. Nachum、G. Tucker 和 S. Levine，“D4RL：深度数据驱动强化学习数据集”，arXiv 预印本 arXiv:2004.07219，2020。