基于因子化子目标扩散的分层实体中心强化学习

Dan Haramati[1], **Carl Qi**[2], **Tal Daniel**[3],**Amy Zhang**[2],**Aviv Tamar**[4†], **George Konidaris**[1†]布朗大学, [2]德克萨斯大学奥斯汀分校, [3]卡内基梅隆大学, [4]以色列理工学院, † 共同指导 dan haramati@brown.edu;carlq@utexas.edu

arXiv:2602.02722v1 [cs.LG] 2 Feb 2026

摘要

我们提出了一种用于离线目标条件强化学习（GCRL）的分层实体中心框架，该框架将子目标分解与因子化结构相结合，以解决包含多个实体的领域中的长时程任务。在复杂环境中实现长时程目标，仍是强化学习（RL）的核心挑战。由于组合复杂度，多实体领域尤其困难。GCRL 有助于在不同目标间实现泛化并利用子目标结构，但在高维观测和组合状态空间下表现不佳，稀疏奖励情形尤为如此。我们采用两级层次结构，由基于价值的目标条件强化学习智能体和因子化的子目标生成条件扩散模型组成。强化学习智能体与子目标生成器分别独立训练，随后通过基于价值函数的选择性子目标生成进行事后组合，使该方法具有模块化特性，并能兼容现有的目标条件强化学习算法。我们对基准任务引入了新的变体，以凸显多实体领域中的挑战；实验表明，在具有稀疏奖励、基于图像的长时域任务中，我们的方法始终能提升底层强化学习智能体的性能，在测试套件中最困难的任务上使成功率提高超过 150%，并能泛化至更长时域和更多实体。轨迹视频见：https://sites.google.com/view/hecrl。

1 引言

规划以在非结构化环境中实现复杂的长时程目标，是强化学习（RL）与序贯决策研究核心中的一项决定性挑战。目标条件强化学习（GCRL）（Kaelbling，1993）提供了一个有用的框架，通过促进目标之间的泛化与知识迁移来训练目标达成智能体，并在与深度强化学习算法结合时展现出良好前景（Schaul 等人，2015）。然而，面对复杂目标分布和高维观测空间时，其有效性仍然有限，尤其是在稀疏奖励条件下（Park 等人，2025年a）。

在目标条件强化学习中利用子目标结构（即轨迹中的中间状态也可被视为目标这一事实），已通过目标重标记（Andrychowicz 等人，2017年）和分层学习（Levy 等人，2019年；Chane-Sane 等人，2021年；Bagaria 等人，2021年）在处理长时程稀疏奖励任务方面取得了显著进展。Park 等人（2023年）提出了一种从单一目标条件价值函数中提取的两层策略层级结构，以解决使用时序差分目标进行学习时不断累积的近似误差。由此得到的算法 HIQL 在离线目标条件强化学习设定中提供了一种简单的层级结构实例化。然而，正如近期的 OGBench 基准（Park 等人，2025年a）所揭示的，离线目标条件强化学习方法（包括 HIQL）在组合状态空间和图像观测方面仍然面临困难。

若要推动目标条件强化学习在现实问题及机器人学等领域中的应用，我们的算法必须能够适应复杂目标分布，以及图像等丰富的高维观测。尤其具有挑战性的是环境包含多个实体的场景，因为状态空间会随着实体数量增加而呈组合式增长。这类领域的例子包括机器人操作物体（Haramati 等，2024）、多机器人路径规划（Shaoul 等，2025）、自动驾驶（Vinitsky 等，2018）和电子游戏（Zambaldi 等人，2019；Delfosse 等，2024）。在这种情况下，可以利用底层的因子化结构，并将其作为归纳偏置，从而显著简化学习过程。

在这项工作中，我们旨在结合子目标层级与因子化结构的优势，以解决多实体领域中的长时程任务。在某些情况下，一种自然的子目标分解是修改部分实体；但正如我们所展示的，HIQL 等最先进的方法并不会鼓励这种稀疏性。我们提出了一种用于离线目标条件强化学习的分层实体中心框架，无需监督即可扩展至图像观测。该框架利用无监督物体中心表征以及多实体环境中的因子化结构，生成实体因子化子目标，即与当前状态相比，实体发生稀疏变化的子目标。

我们采用两级层次结构：低层是一个基于价值的目标条件强化学习智能体，高层则是一个子目标生成条件扩散模型。我们证明，实体中心化扩散所引入的偏差会促进实体因子化子目标的形成，从而简化强化学习策略所面对的子任务。我们的方法兼具模块化与自适应能力：两个层级仅通过训练所用的数据集相关联，并在训练后通过子目标生成过程加以组合；该过程会基于价值函数选择性地生成子目标。这使得该方法能够轻松集成到几乎任何基于价值的目标条件强化学习算法中。

我们展示了能够在组合复杂且观测空间高维的环境中有效实现长时域目标，同时从稀疏奖励中学习的智能体。我们通过对现有基准测试进行创新性改造来测试我们的方法，以突出上述挑战。我们的方法始终能够提升扁平实体中心目标条件强化学习智能体的性能；在我们测试套件中最困难的任务上，成功率提升超过 150%，并且能够泛化至更长的时域和更多的实体。

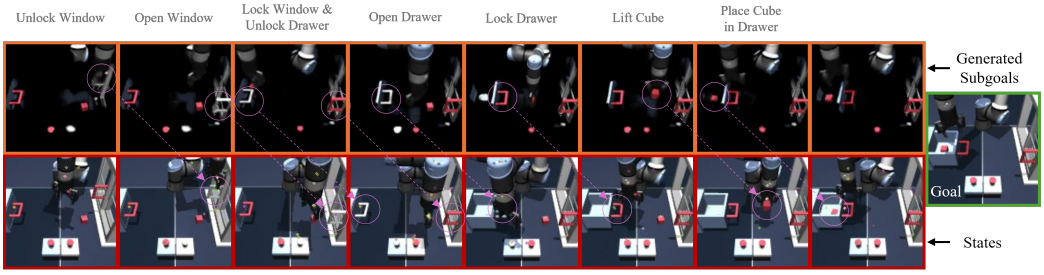

图1：场景**基于图像的因子化子目标**。底行：子目标生成时的环境图像观测。顶行：所生成潜在因子化子目标的重建。最右侧图像：目标图像观测。圆圈和箭头：顶行的圆圈突出显示与其下方图像中的当前观测相比被子目标修改的因子（不包括机械臂）。指向底行圆圈的箭头突出显示由低层强化学习策略操纵的因子，表明这些子目标已达成。子目标重建仅包含由Deep Latent Particles（DLP，Daniel & Tamar（2022年））表示所捕获的场景前景。有关任务B和DLP模型A.3的详细信息，请参见附录。

2 背景与相关工作

目标条件强化学习：目标条件强化学习（Kaelbling，1993）考虑一个马尔可夫决策过程 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mu, p, r)$，其中 $\mathcal{S}$ 表示状态空间，$\mathcal{A}$ 表示动作空间，μ：$\mathcal{P}(\mathcal{S})$ 表示初始状态分布，p：$\mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ 表示环境转移动态，r：$\mathcal{S} \times \mathcal{G} \to \mathbb{R}$ 表示奖励函数，$\mathcal{G}$ 表示目标空间。目标是学习一个策略 π^*：$\mathcal{S} \times \mathcal{G} \to \mathcal{A}$，使其在给定目标分布下最大化期望折扣回报 $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$，$\gamma$ 表示折扣因子。我们的工作聚焦于离线设置，智能体从次优状态-动作轨迹的固定数据集中学习。我们假设 $\mathcal{G} = \mathcal{S}$ 以及稀疏奖励 $r_t(s_t, g) = \mathbf{1}\{s = g\} - 1$，即处于目标时为0，否则为 -1。

分层目标条件强化学习：分层强化学习通过利用序列数据中的时间结构，能够跨越多个时间尺度和抽象层次进行推理（Klissarov 等人，2025年）。在目标条件强化学习中，许多算法通过不同方式实例化子目标层级结构，以应对长时程目标达成问题（Schmidhuber，1991；Dayan 和 Hinton，1992；Kulkarni 等人，2016；Vezhnevets 等人，2017；Nachum 等人，2018；Levy 等人，2019；Chane-Sane 等人，2021；Bagaria 等人，2021）。与我们的工作密切相关的是 HIQL（Park 等人，2023）

它从基于离线数据学习得到的单个价值函数中提取两级策略层次结构。与 HIQL 不同，我们使用扩散模型来建模高层策略；正如我们所展示的，这对于生成高质量且对应于有效状态的子目标至关重要。此外，我们在整个层级结构中融入以实体为中心的结构，以简化组合式多实体状态空间领域中的学习，并促进因子化子目标的生成。

序列决策扩散：扩散模型（Sohl-Dickstein 等人，2015；Ho 等人，2020）通过逆转一个逐渐向干净样本添加噪声的过程来学习生成数据，并训练神经网络执行逐步去噪。该网络通过预测每一步注入的噪声进行优化，并通过逐渐对来自初始噪声分布的样本进行去噪来生成数据。由于能够捕捉复杂的多模态数据分布，扩散模型已广受欢迎。将扩散模型应用于决策的相关工作大致可分为扩散策略（Chi 等人，2023；Hansen-Estruch 等人，2023）和扩散规划器（也称为扩散器）（Janner 等人，2022；Ajay 等人，2023；Lu 等人，2025）：前者对动作（或动作“片段”）的分布进行建模，后者对状态或状态-动作轨迹的分布进行建模。Li 等人（2023年）和 Chen 等人（2024年）提出了分层扩散器，为长时程任务生成子目标轨迹，以此为较低层扩散器提供条件。与上述方法相比，我们的方法采用一个不带引导的高层扩散模型，为目标条件强化学习策略生成单个即时子目标，并利用其价值函数进行测试时子目标选择。我们对 Qi 等人（2025年）提出的以实体为中心的扩散器进行了改造；该扩散器最初用于目标条件行为克隆，现用于生成实体因子化子目标。

以实体为中心的强化学习：以实体为中心的强化学习关注可描述为一组实体的环境，即 $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2 \otimes ... \otimes \mathcal{S}_N$。将这种因子化结构融入状态表示和智能体架构中，可以简化学习（Sanchez-Gonzalez 等人，2018；Zadaianchuk 等人，2022；Sancaktar 等人，2022；Zhou 等人，2022），并促进组合泛化（Lin 等人，2023）。在没有监督的情况下，获得能够近似环境真实状态的图像因子化表示并非易事，也是以往研究的主题。这些研究要么在学习决策模块的同时学习表示（Zambaldi 等人，2019；Watters 等人，2019；Veerapaneni 等人，2020），要么预训练无监督的以物体为中心的表征（Lin 等人，2020；Locatello 等人，2020；Daniel & Tamar，2022），以供下游强化学习使用（Zadaianchuk 等人，2021；Yoon 等人，2023；Haramati 等，2024）。我们提出了一种建立在 Haramati 等人（2024）工作基础上的模块化分层框架，通过因子化子目标扩散，利用稀疏奖励实现长时程目标达成。

扩展背景介绍和相关工作请参见附录 A。

3 层次化实体中心化强化学习

我们的框架——层次化实体中心强化学习（HECRL）——针对多实体领域中阻碍精准价值学习的两个主要问题：长时域上的奖励传播（3.2）和组合式状态复杂度（3.3），这两个问题都会随着实体数量的增加而加剧。我们采用以实体为中心的方法（Haramati 等，2024；Qi 等，2025），并提出一种两级层次结构，由一个基于价值的目标条件强化学习智能体和一个因子化子目标生成扩散模型组成。我们的方法具有模块化和自适应特性，因此兼容各种基于价值的目标条件强化学习算法。此外，即使无法获取底层的因子化环境状态，它也能处理基于图像的领域（例如，基于像素的机器人操作和电子游戏），方法是使用无监督的以物体为中心的表征（例如，我们在实验中使用的 DLPv2（Daniel & Tamar，2024））。

3.1 动机：规避值近似误差的累积

利用学习到的价值函数中的因子化结构，可提升组合状态空间中的性能，但无法完全消除价值近似误差。时序差分（TD）学习（Sutton，1988）会产生随时间步跨度累积的误差。在长时程稀疏奖励任务中，这种误差影响更大，因为奖励必须经过多步折扣时序差分更新才能传播。尽管价值函数仍可用于近似状态与目标之间的距离，但在目标条件策略提取方面可能不那么有效：对于距离给定目标较远的状态，近似误差可能大于相隔一个低层动作的状态之间微小的价值差异。Park 等人（2023年）将此称为

价值信噪比，它会随状态间距离增大而提高。该比率定义了一个“有效半径”，即其中状态的价值能够为策略提取提供清晰学习信号；我们将其称为策略能力半径，为简化记法，记作 R_π^V 或 R。这一量无法先验得知。给定一个离线强化学习策略，我们的方法旨在生成一个即时子目标，该子目标既能被该策略到达（即位于 R_π^V 范围内），又能使策略更接近目标。此过程可以重复进行，直到目标进入策略的可达范围。我们的方法有两个关键方面，使其区别于以往方法：（1）模块化与测试时灵活性——我们不要求对底层强化学习智能体进行任何修改。我们单独训练一个用于生成子目标的条件扩散模型，并在测试时施加基于价值的可达性约束（进一步讨论见图2左侧和附录D.1）。（2）因子化子目标扩散——当数据支持时，我们的实体中心化方法鼓励生成仅对状态因子进行稀疏修改的子目标（实证结果见图2右侧和第4.5节）。当这些因子（或其子集）可独立控制时，只需修改少数状态因子的子目标通常更容易到达，因此更适合作为低层策略的条件。

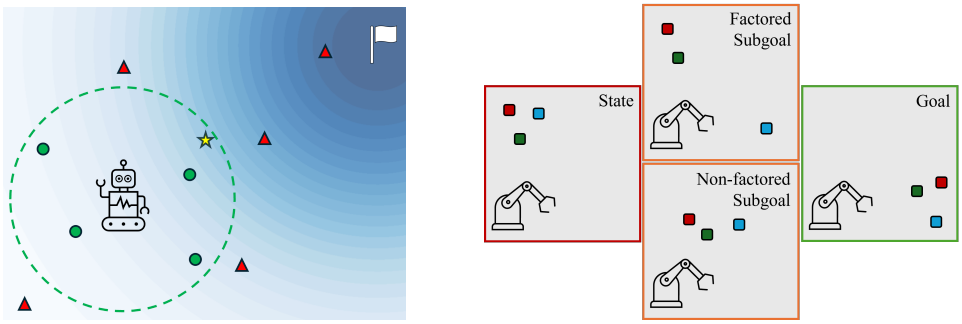

图2：**左：子目标采样示意图**（算法1第2–4行）。机器人 - 智能体，虚线圆 - 能力半径，旗帜 -目标，红色三角形 - 丢弃的子目标，绿色圆形 -过滤后的子目标，黄色星形 - 选定的子目标，背景颜色渐变 - 价值景观。**右：因子化子目标示意图**。因子化子目标便于修改实体的小子集，从而在各因子可独立控制时简化子任务。子目标图像描绘了两个与初始状态距离大致相同的状态，其中上方只需移动蓝色立方体，而下方需要移动全部三个。

3.2 子目标扩散器

我们的方法假设可以访问一个已训练的基于价值的以实体为中心的目标条件强化学习智能体，它由策略 π：$\mathcal{S}\times\mathcal{G}\to\mathcal{A}$ 和价值函数 V：$\mathcal{S}\times\mathcal{G}\to\mathbb{R}$ 组成，以及该智能体训练所用的离线数据集。我们提醒读者，在我们的设定中 $\mathcal{G}=\mathcal{S}$，但为了可读性，我们仍保留 $\mathcal{G}$ 记号。

我们在离线强化学习数据集上训练基于条件扩散的子目标生成器 $\mathcal{D}$：$\mathcal{S}\times\mathcal{G}\to\mathcal{G}$，使其拟合给定目标 g 时，与状态 s 相距至多 K 个时间步的状态分布，其中 g 是从与 s 相同轨迹的未来时间步中均匀采样得到的状态。具体而言，给定一条长度为 T 的离线状态轨迹：(s_0、.....、s_T)，我们 (1) 从 $[0,T-1]$ 到 t 均匀采样一个时间步，以选取状态 $s=s_t$；(2) 从 $[t+1,T]$ 到 t_g 均匀采样一个时间步，以选取目标状态 $g=s_{t_g}$；(3) 将训练子目标设为 $\tilde{g}=s_{min(t+K,t_g)}$；(4) 训练条件扩散去噪器来建模数据集分布 $p(\tilde{g}|s,\ g)$。这一过程可以视为训练一个目标条件行为克隆子目标策略，我们将其称为子目标扩散器。

我们不假设数据集包含目标导向行为，这带来了几项影响：(1)$p(\tilde{g}|s,\ g)$可能具有高度多模态，因此我们选择采用扩散建模。(2) $p(\tilde{g}|s,\ g)$中可能包含与 s 和 g的价值距离各不相同的状态，也就是说，分别为 $V(s,\ \tilde{g})$和 $V(\tilde{g},\ g)$，这使得 K 难以设定，因而无法保证子目标可达（即便我们事先能够获得 R）。(3) 子目标 Diffuser 拟合行为数据，因此无法捕捉子目标最优性的概念。因此，我们提出了一种简单而有效的测试时子目标生成程序，下面将对此进行介绍，并在算法1中予以总结。

我们从子目标扩散器中采样 N 个子目标候选，并根据价值阈值 R筛选其可达性，即保留满足 $V(s,\ \tilde{g})>R$ 的子目标 $\tilde{g}$。随后，我们选择

算法 1子目标生成过程

输入： 当前时间步 t、当前状态 s、当前子目标 g'、目标 g、子目标扩散器 $\mathcal{D}$、价值函数 V、策略能力半径 $\hat{R}$、子目标样本 N、子目标时间步 T_{sg}。
输出： 子目标 $\tilde{g}$

1: **如果** $t\%\, T_{sg} == 0$**则** # 采样新的子目标
2: 采样子目标候选 $\{\tilde{g}_i\}_{i=}^{N}$ $_1 \sim \mathcal{D}(s,\ g)$
3: 筛选可达的子目标 $\{\tilde{g}_j\}_{j=1}^{M} \leftarrow \{\tilde{g}_i \mid V(s, \tilde{g}_i) >$ $\hat{R}\}$ # $M \leq N$
4: 选择最接近目标 $\tilde{g} \leftarrow$ 的子目标 $_{\tilde{g}\in\{\tilde{g}_j\}} V$arg max$\tilde{g}, g$
5: **如果** $V(\tilde{g}, g) \leq V(s, g)$ **那么**
6: $\tilde{g} \leftarrow g$ # 如果 s 比目标 g 更近，则直接前往 g，而不是生成的子目标 $\tilde{g}$
7: **结束如果**
8: **否则** # 保留当前子目标
9: $\tilde{g} \leftarrow g'$
10: **结束if**
11: **返回** $\tilde{g}$

距离目标最近的子目标，即价值 $V(\tilde{g},\ g)$ 最高的子目标（见图 2 左图）。我们使用该子目标 π 执行低层强化学习策略，并运行固定 T_{sg} 个时间步，然后重复子目标生成过程。这可以视为一种结合滚动时域控制或模型预测控制（MPC）的约束采样规划，其中我们的模型由 $\{\mathcal{D}, V\}$ 构成，优化在状态空间 $\mathcal{S}$ 上进行，而非动作空间 $\mathcal{A}$。为帮助收敛，如果目标比所选子目标更接近当前状态，则将其替换为目标（算法1第 5–7 行）。我们对这些子目标施加的约束适用于我们的场景，因为 $-V(s,\ g)$ 可视为状态之间的折扣（非对称）距离指标（Wang 等人，2023），而且我们的方法也兼容其他针对状态的测试时约束。

3.3 生成实体因子化子目标

上一节介绍了一种不依赖状态空间结构的子目标生成与选择机制。当环境状态可以表示为实体集合时，已有研究表明，因子化状态表示和基于集合的策略架构（例如 Transformer（Vaswani等人，2017））能够更好地处理状态空间的组合复杂度，从而提升性能和样本效率，并促进组合泛化（Zhou等人，2022；Haramati 等，2024）。我们进一步利用这种结构来生成实体因子化子目标。除了上述优点，实体级因子化还可以通过便于修改部分实体来简化低层策略的子任务。虽然可以采用明确的方法来约束子目标生成，使其相对于当前状态仅进行稀疏修改，但我们观察到，在适当归纳偏置的作用下，这种特性会自然地从数据中显现出来。在我们的方案中，这种偏置由实体中心扩散（Qi 等，2025）诱导：给定状态实体集合 $s = \{s_m\}_{m=1}^{M}$ 和目标实体集合 $g = \{g_m\}_{m=1}^{M}$，我们的扩散模型会逐步对一组带噪子目标实体 $\tilde{g}^\tau = \{\tilde{g}_m^\tau\}_{m=1}^{M}$进行去噪，其中 M 表示集合大小，τ 表示扩散时间步。重要的是，s、g 和 $\tilde{g}^\tau$ 中的每个实体都是Transformer去噪器（或其他基于集合的架构）的单独输入，并编码了其所属的集合。

如上一节所述，分布 $p(\tilde{g}|s,\ g)$ 包含了多种实现给定目标的方式，其数量会随着实体数量的增加而组合式增长。这使得相关子目标的分布范围广泛且呈多模态。我们选择扩散模型，能够捕捉数据中存在的多个子目标模式，这些模式可能对应于修改不同实体子集的状态。这反过来使我们能够从不同模式中进行采样。相比之下，对这些模式进行加权平均更可能导致对状态的大范围修改（见第4.5节）。正如实验所示，将实体中心表示与基于Transformer的扩散相结合，有助于生成实体因子化子目标。我们认为，这在一定程度上归功于Transformer能够通过注意力机制，有选择地将输入词元（$\{s_m\}_{m=1}^{M}$ 和 $\{g_m\}_{m=1}^{M}$）复制到输出（$\{\tilde{g}_m^\tau\}_{m=1}^{M}$）（Jelassi等人，2024）。此外，以实体为中心的结构使我们能够训练规模相对较小的扩散模型，去噪步数最少可达10步，从而降低实时控制的计算负担。

4 实验

实验旨在研究：（1）我们的方法在处理涉及多个实体的长时域目标条件任务时的有效性（第4.3节）；（2）我们的设计选择对性能的影响（第4.4节）；（3）我们的方法生成的子目标的质量（第4.5节）；（4）我们方法的组合泛化能力（第4.6节）。我们在第4.1节介绍评估套件，并在第4.2节说明实现细节和基线方法。我们提供了数据集、检查点和代码，以便复现本文结果：https://github.com/DanHrmti/HECRL。

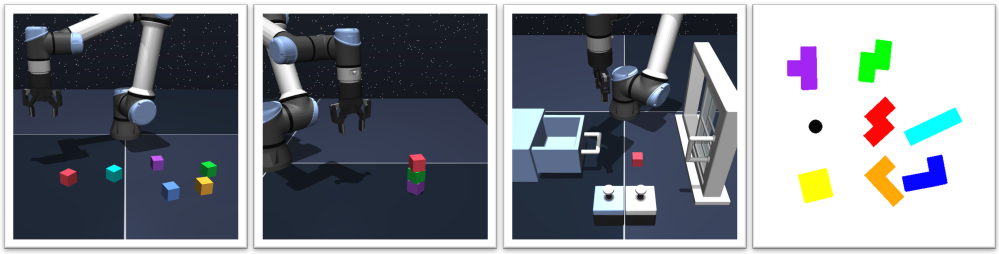

图 3：**环境套件**。PPP-Cube：**抓**取、**放**置和**推**动立方体至目标位置。Stack-Cube：抓取-放置立方体堆叠。场景：将物体（抽屉、窗户、按钮和立方体）操作至目标配置。Push-Tetris：将方块推动至目标位置和朝向。

4.1 环境、任务和数据集

我们的实验结果聚焦于多物体操作，使用一套改编自先前基准测试的环境、任务和数据集，以凸显感知以及处理组合状态空间方面的挑战。这些环境的可视化见图 3。我们这套环境的大部分基于 OGBench（Park 等人，2025年a）的操作环境，具体为立方体和场景环境。这些环境包含 UR5e 机械臂，并涉及桌面物体操作。我们提高了图像分辨率，增加了多视角感知，并移除了简化视觉成分，例如机械臂透明度和末端执行器着色。

任务：(1) PPP-Cube：基于 OGBench 立方体 环境，智能体需要在随机初始化的状态与目标位置之间操作立方体。每个回合开始时，从固定集合中为每个立方体随机分配互不相同的颜色。配套数据集包含多样的智能体与物体的交互，包括**抓**取-**放**置和**推**动操作。图像观测包含正面和侧面两个视图。为了将算法方面与任何特定的以物体为中心的图像表示隔离开来，我们支持基于状态的实体中心观测，以模拟一种“完美”的因子化表示。(2) Stack-Cube：除数据集仅包含高概率堆叠的抓取与放置操作、且智能体仅针对堆叠进行评估外，其余与 PPP-Cube 相同。(3) 场景：我们原样采用 OGBench 视觉场景 环境，并对部分测试任务进行修改，以避免因我们更真实的视觉设置而在目标描述中产生的歧义（见附录 B）。该环境从单一视图进行观测。(4) Push-Tetris：我们将 Chi 等人（2023）提出的 *Push-T* 环境改造为对俄罗斯方块状方块的多物体操作。每个回合开始时随机采样不同的方块类型。每种方块类型具有不同的颜色，且该颜色在各回合之间保持固定。智能体需要将方块推动到包含位置和方向的目标配置。数据集通过随机策略采集：该策略被限制在物体周围的固定半径内，并按固定时间间隔进行采样，从而导致高度次优行为。扩展的环境细节见附录 B。

环境特征：我们认为，PPP-Cube 和 Stack-Cube 是我们测试套件中最具挑战性的任务[1]，，因为它们要求从真实的图像观测中学习在组合状态空间内进行长时程三维操作。Scene 在感知方面面临类似的挑战，且包含更多物体类型（以及相应的操作能力），但有效状态配置少得多，因此其组合复杂度显著较低。Push-Tetris 则

[1]参见附录 B.1 的文献综述，其中指出：在我们所考虑的数据条件下，对于包含超过 2 个物体的这些环境，在本工作之前尚未取得非平凡的性能。

视觉上更简单，且局限于二维平面，但要求在组合状态空间中对不同物体进行细粒度控制。

评估：训练数据集包含 3 个物体（不包括场景），最多包含 $3M$ 次状态转移。我们为智能体设置固定数量的梯度更新，并在所有实验和指标中报告 4 个随机种子下最终检查点的均值和标准差。最佳结果（在一个标准差范围内）以粗体标出。与 OGBench 不同，我们在测试时不会在达到目标后终止，这样智能体就不能仅仅偶然碰到目标，而必须到达目标并稳定地维持它。这对于 Stack-Cube 之类的任务至关重要，因为立方体应保持堆叠状态；对于场景任务也同样重要，因为目标配置的有效数量足以让仅通过探索这些配置、而不依赖目标描述的策略取得非平凡的性能。

4.2 实现与基线方法

我们精心选择了基线方法，以对比本方法的不同方面和设计选择（见附录表 9），同时确保比较公平。我们考虑的强化学习智能体基于 IQL 的目标条件变体（Kostrikov 等人，2022年）。

EC-SGIQL（我们的方法）：我们在使用 IQL 训练的 ECRL（Haramati 等，2024）智能体基础上实现了我们的方法。对于子目标扩散器，我们改造了 EC-Diffuser（Qi 等，2025），使其能够根据以实体为中心的状态和目标生成子目标。我们将这一方法实例称为实体中心子目标 IQL（EC-SGIQL）。

EC-IQL：对应于一个通过将实体交互Transformer（EIT）架构与 IQL 相结合而适配至离线设置的 ECRL（Haramati 等，2024）智能体。该基线代表一种以实体为中心的非层次化方法，相当于不进行子目标生成的我们的方法。**EC-Diffuser**（Qi 等，2025）：一种以实体为中心的扩散式行为克隆方法。

HIQL：一个基于 HIQL（Park 等人，2023）的智能体。
IQL：一个基于 IQL（Kostrikov 等人，2022年）的智能体。
所有以实体为中心的方法（EC- 前缀）均使用以实体为中心的状态观测进行训练（详情见附录 B），或使用通过预训练的 DLPv2（Daniel & Tamar，2024）提取的潜在图像表示（DLP 概述见附录 A.3）。标准智能体使用单向量状态观测进行训练，或使用通过预训练的 VQ-VAE（Van Den Oord 等，2017年）提取的潜在图像表示。图像表示是在离线强化学习数据集上预训练的，每项任务各使用一个数据集（重建可视化见附录 D.6）。实现细节详见附录 C。

4.3 长时域操作

表1总结了所有方法在我们环境套件上的性能。对于 Push-Tetris，我们报告物体平均状态-目标像素覆盖率（即重叠度）；其他环境则报告成功率。除两种表现相当的情况外，我们的方法显著优于所有基线方法。值得注意的是，EC-SGIQL 使用与 EC-IQL 相同的低层策略和价值函数，但性能始终更佳。在图像输入的 PPP-Cube 任务中（这是我们环境套件中最具挑战性的任务），其成功率提升了 150% 以上。EC-IQL 是性能第二好的方法，这凸显了结构在这些领域中的重要性。图1和图4展示了在基于图像的环境中，我们的方法遵循 Diffuser 生成的子目标执行的轨迹。有关进一步讨论和性能结果，请参阅附录D.2；轨迹视频请见我们的网站。

表1：**长时程操作性能**。除 Push-Tetris 外，所有数值均为成功率；对于 Push-Tetris，我们报告状态-目标像素覆盖率。基线详情见第4.2节。与标准差范围内的最佳结果以**粗体**标出。

环境	EC-SGIQL	EC-IQL	EC-Diffuser	HIQL	IQL
PPP-Cube (状态)	**82.5** ± **3.1**	51.5 ± 4.4	44.8 ± 6.7	48.3 ± 7.3	34.3 ± 4.9
Stack-Cube (状态)	**43.5** ± **1.9**	29.0 ± 2.9	**43.8** ± **9.2**	0.0 ± 0.0	19.3 ± 3.0
PPP-Cube (图像)	**64.3** ± **4.9**	25.0 ± 5.7	0。3 ± 0.5	0.0 ± 0.0	0.0 ± 0.0
场景 (图像)	**61.5** ± **5.9**	**53.0** ± **5.5**	3.3 ± 2.5	8.3 ± 1.3	17.5 ± 2.7
Push-Tetris (图像)	**61.4** ± **3.3**	31.6 ± 1.3	7.9 ± 0.5	5.2 ± 0.8	3.4 ± 0.8

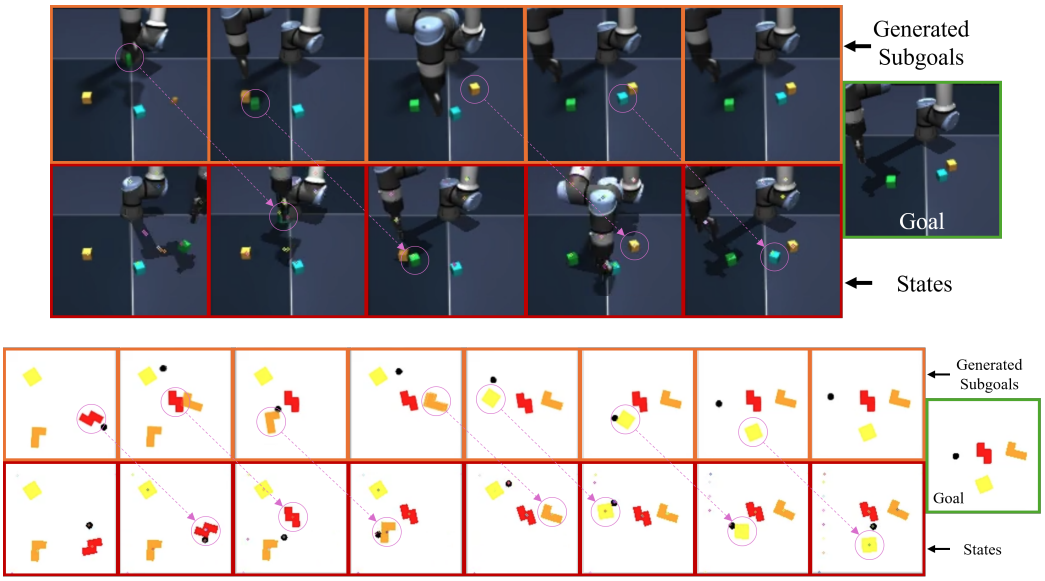

图 4：PPP-Cube 和 Push-Tetris 的**基于图像的因子化子目标**。底行：生成子目标时的环境图像观测。图中标出了潜在粒子的位置。顶行：生成的潜在子目标的 DLP 重建。最右侧图像：目标图像观测。圆圈和箭头：顶行图像中的圆圈标出与其下方图像中的当前观测相比被子目标修改的因子（不包括智能体）。指向底行圆圈的箭头标出由低层强化学习策略操控的因子，表明子目标已实现。

4.4 消融研究

表2展示了针对子目标生成过程以下方面进行消融实验的结果：(1) 使用价值函数选择子目标（随机采样）；(2) 使用价值函数筛选子目标（算法 1 第3行）（最大价值）；以及 (3) 通过训练一个基于确定性 Transformer 的子目标生成器，并采用优势加权回归（*AWR*，Peng 等人，2019）目标函数（如 HIQL 中所用），来考察扩散的作用。使用以实体为中心的潜在图像表示进行 AWR 时需要特别处理，因为潜在实体没有顺序。我们将标准回归损失调整为基于倒角距离的损失（详情见附录 C）。

我们的方法在成功率和时间步效率方面始终取得最佳结果。最大价值变体在部分任务上的表现不相上下，这表明在这些情况下，策略能力半径可能大于生成子目标的半径。我们观察到的另一趋势是，在成功率相近时，随机采样 和*AWR* 变体需要更多时间步才能达到目标。从 Diffuser 中随机采样时，样本并不总能直接推动任务朝目标前进，因此这一结果符合预期。*AWR*变体的整体性能最低，我们认为这是由子目标质量所致。使用加权回归拟合（潜在）观测生成模型，会得到观测的加权平均值，而这些平均值通常并不对应有效状态。我们将在下一节进一步探讨这一问题。

4.5 子目标质量的衡量

我们推测，与基线方法和消融实验相比，我们的方法表现出色，部分原因在于：当数据和领域条件允许时，它会鼓励生成更简单的子目标，这些子目标相较于当前状态，对因子的改动较少。我们在基于状态的立方体 环境中测量这种稀疏性，并在表 3 中报告被修改实体的平均数量。使用 AWR 训练子目标策略，大多数情况下会改动全部 3 个立方体；而我们的子目标扩散器平均只改动接近 1 个立方体。我们认为，这得益于扩散模型能够从不同的子目标模式中生成样本，而确定性策略会生成这些模式的加权平均；此外，子目标扩散器的因子化结构也发挥了作用。

表 2：**子目标生成消融实验性能**。我们的方法：EC-SGIQL。最大价值：我们的方法，但不筛选子目标候选项。随机采样：从 Diffuser 中随机采样一个子目标。*AWR*：将 Diffuser 替换为使用 AWR 训练的确定性 Transformer。对于 Push-Tetris，我们报告像素覆盖率（上）和整个回合内像素倒角距离之和（下）。其余任务报告成功率（上）和负回报（下），后者对应于完成任务或因达到评估时间步上限而终止所用的平均时间步数。最佳结果（标准差范围内）以**粗体**标出。

环境	Ours	最大价值	随机采样	AWR
PPP-Cube （状态）	**82.5** ± 3.1	76.3 ± 1.0	**73.0** ± 7.4	67.8 ± 5.7
	422 ± 60	**464** ± 35	759 ± 23	773 ± 46
Stack-Cube (状态)	**43。5** ± 1.9	**35.5** ± 7.5	28.8 ± 5.5	9.0 ± 4.2
	706 ± 3	771 ± 54	904 ± 21	985 ± 6
PPP-Cube (图像)	**64.3** ± 4.9	46.3 ± 2.1	**57.3** ± 8.7	55.5 ± 3.4
	587 ± 18	709 ± 7	750 ± 42	735 ± 37
场景 (图像)	**61.5** ± 5.9	**61.0** ± 6.7	41.0 ± 6.2	44.5 ± 6.0
	597 ± 26	**589** ± 43	843 ± 5	719 ± 45
Push-Tetris (图像)	**61。4** ± 3.3	**59.5** ± 3.4	**58.6** ± 4.5	**56.1** ± 5.6
	510 ± 12	**505** ± 32	622 ± 30	606 ± 40

表 3：**生成的子目标中相较于输入状态发生变化的实体数量（不包括智能体）**。*EC-Diffusion*：我们的子目标 Diffuser；*EC-AWR*：在以实体为中心的状态上使用 AWR 训练的 Transformer；*AWR*：在单向量状态上使用 AWR 训练的多层感知机。如果某个实体的位置变化超过阈值距离，则认为该实体发生了变化。环境中包含 3 个立方体和机械臂。数值为对随机抽取的 400 个初始状态和目标取平均所得。最佳结果（误差范围为一个标准差）以**粗体**标出。

环境	EC-Diffusion（我们的方法）	EC-AWR	AWR
PPP-Cube (状态)	**1.36** ± 0.01	2.96 ± 0.01	2.98 ± 0.00
Stack-Cube （状态）	**1.04** ± 0.01	2。82 ± 0.02	2.98 ± 0.01

由于在基于图像的环境中测量子目标稀疏性并不直接，我们通过使用 DLP 解码器重建子目标扩散器中的潜在子目标粒子，提供定性观察结果：相关结果见图 1 和图 4；AWR 变体的结果见图 6、图 7 和图 8（附录）。Diffuser子目标通常由输入状态和目标图像组合而成，并且相较于状态只包含稀疏变化。生成的子目标并不完美，偶尔会包含同一个实体两次。在这种情况下，这些实体通常分别是两个输入中的精确副本，即一个来自状态，另一个来自目标。这支持了我们的假设：Transformer 会选择性地将输入实体复制到输出中。我们将这类错误归因于 DLP 表征和扩散目标函数的性质，而使用更大的扩散 Transformer 可能可以缓解这些问题。AWR 子目标通常会在一个单个子目标中包含多个代表不同未来的实体副本，从而为低层强化学习策略提供含义不明确的目标。关于这一点的进一步讨论，请参见附录 D.5。

4.6 泛化

引入实体因子化结构可通过促进组合泛化来降低状态空间的组合复杂度（Lin 等，2023）。只要个体因子具有足够的覆盖率，学习所需的状态空间覆盖率就会降低。实体级状态泛化（例如，对物体位置和颜色的新颖组合进行泛化）通过样本效率来衡量。事实上，我们观察到，与非结构化方法相比，以实体为中心的方法只需少得多的梯度更新便可达到峰值性能（后者能够取得非平凡性能时）。通过改变状态实体和/或目标实体的数量，可以测试系统全局状态层面的组合性。我们在这些情形下测试了方法的泛化能力，并在表 4 中报告了性能（包括 Push-Tetris 的结果见附录表 14）。我们的方法

展现出非平凡的泛化能力，但其性能会随着物体数量的增加而下降。我们超越了扁平实体中心型智能体的性能，表明子目标保持了足够高的质量，能够引导强化学习策略。这些结果表明，我们的方法具有通过课程学习或离线到在线微调来适应不断增加的实体数量的扩展潜力，这将留待未来研究。

表 4：**零样本组合泛化**。所有值均为成功率。* 表示每个区段中用于在其他变体上评估零样本能力的训练变体。最佳结果（误差在一个标准差以内）以**粗体**标出。

变体	EC-SGIQL	EC-IQL	EC-Diffuser
2-Cubes-Stack (状态)	69.8 ± 7.0	45.8 ± 12.3	**87.8 ± 5.4**
3-Cubes-Stack* (状态)	**43.5 ± 1.9**	29.0 ± 2.9	**43.8 ± 9.2**
4立方体堆叠-2 (状态)	**28.8 ± 3.9**	17.3 ± 2.5	1.5 ± 1.3
PPP-1立方体 (状态)	**97.0 ± 1.8**	**92。3 ± 8.9**	85.0 ± 9.2
PPP-2立方体 (状态)	**94.0 ± 3.4**	75.0 ± 3.8	76.5 ± 7.6
PPP-3-立方体* (状态)	**82.5 ± 3.1**	51.5 ± 4.4	44.8 ± 6.7
PPP-4-立方体 (状态)	**65.3 ± 3.5**	31.8 ± 4.0	42.8 ± 11.4
PPP-5-立方体 (状态)	**49.0 ± 8.5**	19.3 ± 5.9	23.8 ± 6.7
PPP-6-立方体 (状态)	**25.7 ± 1.5**	10.5 ± 6.1	8.8 ± 3.5
PPP-1立方体 (图像)	**94。5 ± 2.9**	**91.5 ± 1.3**	3.8 ± 1.3
PPP-2立方体 (图像)	**77.0 ± 3.6**	47.8 ± 6.2	0.0 ± 0.0
PPP-3-立方体* (图像)	**64.3 ± 6.0**	25.0 ± 5.7	0.3 ± 0.5
PPP-4-立方体 (图像)	**38.3 ± 5.7**	11.5 ± 4.7	0.0 ± 0.0
PPP-5-立方体 (图像)	**19.3 ± 6.2**	4.0 ± 2.0	0.0 ± 0.0
PPP-6-立方体 (图像)	**9.5 ± 1.3**	1.5 ± 1.3	0.0 ± 0.0

5 结论

我们提出了分层实体中心强化学习（HECRL），这是一种离线目标条件强化学习框架，它将子目标层级与因子化结构相结合，用于解决多实体领域中的长时程任务，并可扩展至图像观测。我们将方法设计得简单、灵活且模块化，使其能够兼容各种基于价值的目标条件强化学习算法和测试时子目标约束。实验结果表明，我们的因子化子目标扩散器与简单的基于价值的选择过程相结合，能够生成高质量的子目标，以引导低层实体中心强化学习智能体，并持续提升其性能。我们研究了影响该方法性能的因素，发现与常用的加权回归目标相比，实体中心扩散所引入的偏差会促使生成的子目标相较于当前状态仅作稀疏修改。最后，我们展示了非平凡的零样本泛化性能：随着实体数量增加（从而时域变长），这种性能仍得以保持，表明该方法有望扩展到包含更多实体的环境。

局限性与未来工作：我们的方法假设价值函数能够为低层策略提供足够的信号，使其具备不可忽略的能力半径，同时也能为子目标生成器提供指导。这一假设在我们的挑战性领域中成立，但可能会限制其在其他领域中的适用性。尽管如此，我们框架的模块化特性有助于基于平面价值的目标条件强化学习算法的进展进行改进。子目标扩散器使用距离当前状态固定 K 步以内的子目标进行训练。尽管得益于测试时的子目标筛选机制（见附录 D.4），我们的方法对这一超参数更加稳健，但未来的工作可以探索如何从数据中自动推断 K 。最后，将我们的方法应用于具有图像观测的领域依赖于良好的因子化状态估计器。我们展示了在无需监督的情况下，这在我们的模拟领域中是可行的。因此，将我们的方法扩展到真实世界自然场景的潜力，与无监督以对象为中心的表征学习的发展密切相关（Seitzer 等，2023年；Zadaianchuk 等，2023年；Daniel 与 Tamar，2024年）。

伦理声明

本工作遵循 ICLR 道德准则。我们的工作推进了离线目标条件强化学习（目标条件强化学习）的方法。本文实验在模拟环境中进行，不涉及人类受试者，也不存在隐私问题；我们预计本工作不会带来任何伦理或社会风险。

可复现性声明

我们致力于确保本研究能够被完整复现。我们在论文正文和附录中详细说明了实现和评估方法。我们公开了代码、检查点和数据集，以便复现本文的所有结果，并促进未来研究：https://github.com/DanHrmti/HECRL。

致谢

本研究得到了美国海军研究办公室的资助，资助编号为 N00014-22-1-2592。
AZ和CQ获得美国国家科学基金会2340651、美国国家科学基金会2402650、美国国防高级研究计划局HR00112490431以及美国陆军研究办公室W911NF-24-1-0193的资助。AT获得欧盟（欧洲研究委员会，Bayes-RL，项目编号101041250）的资助。然而，所表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会执行机构的观点。欧盟和资助机构均不对此承担责任。

参考文献

Anurag Ajay、Yilun Du、AbhiGupta、Joshua B. Tenenbaum、TommiS. Jaakkola和Pulkit Agrawal。决策是否只需要条件生成建模？见第十一届学习表征国际会议，2023年。网址：https://openreview.net/ forum?id=sP1fo2K9DFG。

Marcin Andrychowicz、Filip Wolski、Alex Ray、Jonas Schneider、Rachel Fong、Peter Welinder、Bob McGrew、Josh Tobin、OpenAI Pieter Abbeel和Wojciech Zaremba。事后回顾经验回放。神经信息处理系统进展，30，2017年。

Akhil Bagaria、Jason K Senthil 和 George Konidaris。利用深度技能图进行探索与规划的技能发现。发表于国际机器学习会议，第 521–531 页。PMLR，2021年。

Elliot Chane-Sane、Cordelia Schmid 和 Ivan Laptev。利用想象子目标进行目标条件强化学习。发表于国际机器学习会议，第 1430–1440 页。PMLR，2021年。

MichaelChang、Alyssa Li Dayan、Franziska Meier、ThomasL.Griffiths、Sergey Levine 和 Amy Zhang。用于物体重排中组合泛化的层次抽象。发表于第十一届国际学习表征会议，2023年。网址：https:// openreview.net/forum?id=fGG6vHp3W9W。

Chang Chen、Fei Deng、Kenji Kawaguchi、Caglar Gulcehre 和 Sungjin Ahn。基于扩散的简单层次规划。发表于第十二届国际学习表征会议，2024年。网址：https://openreview.net/forum?id=kXHEBK9uAY。

Cheng Chi、Siyuan Feng、Yilun Du、Zhenjia Xu、Eric Cousineau、Benjamin Burchfiel 和 Shuran Song。《扩散策略：通过动作扩散进行视觉运动策略学习》。载于 机器人学：科学与系统（*RSS*）会议论文集，2023年。

Tal Daniel 与 Aviv Tamar。基于深度潜在粒子的无监督图像表征学习。载于国际机器学习会议，第4644–4665页。PMLR，2022年。

Tal Daniel 与 Aviv Tamar。DDLP：基于深度动态潜在粒子的无监督以物体为中心的视频预测。机器学习研究汇刊，2024年。ISSN 2835-8856。URL https://openreview.net/forum?id=Wqn8zirthg。

Peter Dayan 和 Geoffrey E Hinton。封建式强化学习。神经信息处理系统进展，5，1992年。Quentin Delfosse、Jannis Bl¨uml、Bjarne Gregori、Sebastian Szt wiertnia 和 Kristian Kersting。OCAtari：以物体为中心的 Atari 2600 强化学习环境。载于强化学习会议，2024年。URL https://openreview.net/forum?id=24OiW9d5k8。Stefano Ferraro、Pietro Mazzaglia、Tim Verbelen 和 Bart Dhoedt。Focus：面向机器人操作的以物体为中心的世界模型。神经机器人学前沿，19:1585386，2025年。Scott Fujimoto 和 Shixiang Shane Gu。离线强化学习的极简方法。神经信息处理系统进展，34:20132–20145，2021年。Philippe Hansen-Estruch、Ilya Kostrikov、Michael Janner、Jakub Grudzien Kuba 和 Sergey Levine。IDQL：将隐式 Q 学习作为带扩散策略的演员-评论家方法。*arXiv* 预印本 *arXiv:2304.10573*，2023年。

Dan Haramati、Tal Daniel 和 Aviv Tamar。基于像素的物体操作的实体中心化强化学习。载于第十二届国际学习表征会议，2024年。URL https://openreview.net/forum?id=uDxeSZ1wdI。

Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020年。Shengyi Huang、Rousslan Fernand Julien Dossa、Chang Ye、Jeff Braga、Dipam Chakraborty、Ki- nal Mehta 和 Jo˜ao G.M. Ara´ujo。Cleanrl：深度强化学习算法的高质量单文件实现。机器学习研究杂志，23(274):1–18，2022年。URL http://jmlr.org/papers/v23/21-1342.html。Michael Janner、Yilun Du、Joshua Tenenbaum 和 Sergey Levine。基于扩散的灵活行为合成规划。载于国际机器学习会议，2022年。Samy Jelassi、David Brandfonbrener、Sham M Kakade 和 Eran Malach。重复跟我说：在复制任务上，Transformer 优于状态空间模型。载于国际机器学习会议，第 21502–21521 页。PMLR，2024年。Jindong Jiang*、Sepehr Janghorbani*、Gerard De Melo 和 Sungjin Ahn。SCALOR：具有可扩展物体表征的生成式世界模型。载于国际学习表征会议，2020年。URL https://openreview.net/forum?id=SJxrKgStDH。

Leslie Pack Kaelbling。《学习如何实现目标》。载于*IJCAI*，第 2 卷，第 1094–8 页，1993 年。

Diederik P Kingma 和 Max Welling。自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013年。Martin Klissarov、Akhil Bagaria、Ziyan Luo、George Konidaris、Doina Precup 和 Marlos C Machado。《发现时间结构：分层强化学习概述》。*arXiv* 预印本 *arXiv:2506.14045*，2025年。Ilya Kostrikov、Ashvin Nair 和 Sergey Levine。《基于隐式 Q 学习的离线强化学习》。发表于国际学习表征会议，2022年。网址：https:// openreview.net/forum?id=68n2s9ZJWF8。Tejas D Kulkarni、Karthik Narasimhan、Ardavan Saeedi 和 Josh Tenenbaum。《层次化深度强化学习：整合时间抽象与内在动机》。发表于神经信息处理系统进展，第29卷，2016年。Andrew Levy、George Konidaris、Robert Platt 和 Kate Saenko。《利用事后回顾学习多层级层次结构》。发表于国际学习表征会议，2019年。

Qiyang Li、Zhiyuan Zhou 和 Sergey Levine。《通过动作分块实现强化学习》。发表于 *ICML 2025* 的 *The Exploration in AI Today Workshop*，2025年。Richard Li、Allan Jabri、Trevor Darrell 和 Pulkit Agrawal。《利用关系强化学习实现实用的多物体操作》。发表于*2020*年 *IEEE* 机器人学与自动化国际会议（*ICRA*），第4051–4058页。IEEE，2020年。Wenhao Li、Xiangfeng Wang、Bo Jin 和 Hongyuan Zha。《用于离线决策的层次扩散》。发表于国际机器学习会议，第20035–20064页。PMLR，2023年。Baihan Lin、Djallel Bouneffouf 和 Irina Rish。《应用中的组合泛化综述》。*arXiv* 预印本 *arXiv:2302.01067*，2023年。Zhixuan Lin、Yi-Fu Wu、Skand Vishwanath Peri、Weihao Sun、Gautam Singh、Fei Deng、Jindong Jiang 和 Sungjin Ahn。《Space：通过空间注意力与分解实现无监督面向对象场景表示》。发表于国际学习表征会议，2020年。网址：https://openreview.net/forum?id=rkl03ySYDH。Francesco Locatello、Dirk Weissenborn、Thomas Unterthiner、Aravindh Mahendran、Georg Heigold、Jakob Uszkoreit、Alexey Dosovitskiy 和 Thomas Kipf。《基于槽位注意力的以物体为中心的学习》。发表于神经信息处理系统进展，第33卷，第11525–11538页，2020年。Haofei Lu、Dongqi Han、Yifei Shen 和 Dongsheng Li。《什么样的扩散规划器适用于决策？》。发表于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=7BQkXXM8Fy。

Davide Mambelli、Frederik Tr¨auble、Stefan Bauer、Bernhard Sch¨olkopf 和 Francesco Locatello。基于线性关系网络的组合式多对象强化学习。*arXiv* 预印本 *arXiv:2201.13388*，2022年。

Aditya Mohan、Amy Zhang 和 Marius Lindauer。《深度强化学习中的结构：综述与开放问题》。人工智能研究杂志，79：1167–1236，2024年。Malte Mosbach、Jan Niklas Ewertz、Angel Villar-Corrales 和 Sven Behnke。《Sold：用于从像素学习关系操作的槽位对象中心潜在动力学模型》。*arXiv* 预印本 *arXiv:2410.08822*，2024年。Ofir Nachum、Shixiang Shane Gu、Honglak Lee 和 Sergey Levine。《数据高效的层次强化学习》。神经信息处理系统进展，31，2018年。Seohong Park、Dibya Ghosh、Benjamin Eysenbach 和 Sergey Levine。《HIQL：将潜在状态作为动作的离线目标条件强化学习》。神经信息处理系统进展，36：34866–34891，2023年。Seohong Park、Kevin Frans、Benjamin Eysenbach 和 Sergey Levine。《OGBench：离线目标条件强化学习基准测试》。载于国际学习表征会议（*ICLR*），2025a。Seohong Park、Kevin Frans、Deepinder Mann、Benjamin Eysenbach、Aviral Kumar 和 Sergey Levine。《时间范围缩减使强化学习具有可扩展性》。*arXiv* 预印本 *arXiv:2506.04168*，2025b。Xue Bin Peng、Aviral Kumar、Grace Zhang 和 Sergey Levine。《优势加权回归：简单且可扩展的离策略强化学习》。*arXiv* 预印本 *arXiv:1910.00177*，2019年。Carl Qi、Dan Haramati、Tal Daniel、Aviv Tamar 和 Amy Zhang。《EC-Diffuser：通过以实体为中心的行为生成实现多物体操作》。载于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=o3pJU5QCtv。Cansu Sancaktar、Sebastian Blaes 和 Georg Martius。《通过结构化世界模型进行好奇探索，实现零样本对象操作》。神经信息处理系统进展，35：24170–24183，2022年。

Alvaro Sanchez-Gonzalez、Nicolas Heess、Jost Tobias Springenberg、Josh Merel、Martin Riedmiller、Raia Hadsell 和 Peter Battaglia。《作为可学习物理引擎用于推断与控制的图网络》。载于国际机器学习会议，第4470–4479页。PMLR，2018年。

Tom Schaul、Daniel Horgan、Karol Gregor 和 David Silver。通用价值函数近似器。载于国际机器学习会议，第 1312–1320 页。PMLR，2015年。

J¨urgen Schmidhuber。学习为动作序列生成子目标。载于人工神经网络，第 967–972 页，1991 年。

Maximilian Seitzer、Max Horn、Andrii Zadaianchuk、Dominik Zietlow、Tianjun Xiao、Carl-Johann Simon-Gabriel、Tong He、Zheng Zhang、Bernhard Sch¨olkopf、Thomas Brox 和 Francesco Lo-catello。弥合通向真实世界的以对象为中心的学习之间的鸿沟。载于第十一届国际学习表征会议，2023年。网址：https://openreview.net/forum? id=b9tUk-f aG.

Yorai Shaoul、Itamar Mishani、Shivam Vats、Jiaoyang Li 和 Maxim Likhachev。基于扩散模型的多机器人运动规划。载于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=AUCYptvAf3。

Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。利用非平衡热力学进行深度无监督学习。载于国际机器学习会议，第 2256–2265 页。PMLR，2015 年。

Richard S Sutton。通过时间差分方法学习预测。机器学习，3：9–44，1988 年。

Emanuel Todorov、Tom Erez 和 Yuval Tassa。MuJoco：用于基于模型的控制的物理引擎。载于*2012*年*IEEE/RSJ*国际智能机器人与系统会议，第 5026–5033 页。IEEE，2012 年。

Aaron Van Den Oord、Oriol Vinyals 等。神经离散表征学习。神经信息处理系统进展，第 30 卷，2017年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。神经信息处理系统进展，第 30 卷，2017年。

Rishi Veerapaneni、John D Co-Reyes、Michael Chang、Michael Janner、Chelsea Finn、Jiajun Wu、Joshua Tenenbaum 和 Sergey Levine。基于视觉模型的强化学习中的实体抽象。载于《机器人学习会议》，第 1439–1456 页。PMLR，2020年。

Alexander Sasha Vezhnevets、Simon Osindero、Tom Schaul、Nicolas Heess、Max Jaderberg、David Silver 和 Koray Kavukcuoglu。用于分层强化学习的封建网络。载于《国际机器学习会议》，第 3540–3549 页。PMLR，2017年。

Eugene Vinitsky、Aboudy Kreidieh、Luc Le Flem、Nishant Kheterpal、Kathy Jang、Cathy Wu、Fangyu Wu、Richard Liaw、Eric Liang 和 Alexandre M Bayen。混合自主交通中的强化学习基准测试。载于《机器人学习会议》，第 399–409 页。PMLR，2018年。

Tongzhou Wang、Antonio Torralba、PhillipIsola 和 Amy Zhang。通过准度量学习实现最优目标到达强化学习。载于《国际机器学习会议》，第 36411–36430 页。PMLR，2023年。

Nicholas Watters、Loic Matthey、Matko Bosnjak、Christopher P Burgess 和 Alexander Lerchner。Cobra：通过无监督物体发现和好奇心驱动的探索实现数据高效的基于模型强化学习。*arXiv* 预印本 *arXiv:1905.09275*，2019年。

Jaesik Yoon、Yi-Fu Wu、Heechul Bae 和 Sungjin Ahn。关于为强化学习预训练以物体为中心的表征的研究。载于国际机器学习会议（国际机器学习会议）*2023*年，第 1–28 页，2023年。

Andrii Zadaianchuk、Maximilian Seitzer 和 Georg Martius。利用以物体为中心的表征进行自监督视觉强化学习。载于国际学习表征会议，2021年。网址：https://openreview.net/forum?id=xppLmXCbOw1。

Andrii Zadaianchuk、Georg Martius 和 Fanny Yang。利用可独立控制的子目标进行自监督强化学习。载于机器人学习会议，第 384–394 页。PMLR，2022年。

Andrii Zadaianchuk、Maximilian Seitzer 和 GeorgMartius。通过预测时间特征相似性，对真实世界视频进行以物体为中心的学习。神经信息处理系统进展，36：61514–61545，2023年。

Vinicius Zambaldi、David Raposo、Adam Santoro、Victor Bapst、Yujia Li、Igor Babuschkin、Karl Tuyls、David Reichert、Timothy Lillicrap、Edward Lockhart、Murray Shanahan、Victoria Langston、Razvan Pascanu、Matthew Botvinick、Oriol Vinyals 和 Peter Battaglia。具有关系归纳偏置的深度强化学习。载于国际学习表征会议，2019年。URL https://openreview.net/forum?id=HkxaFoC9KQ。

Weipu Zhang、Adam Jelley、Trevor McInroe 和 Amos Storkey。物体至关重要：以物体为中心的世界模型能够提升视觉复杂环境中的强化学习效果。*arXiv* 预印本 *arXiv:2501.16443*，2025年。

Linfeng Zhao、Lingzhi Kong、Robin Walters 和 Lawson LS Wong。迈向面向对象的世界建模中的组合泛化。载于国际机器学习会议，第 26841–26864 页。PMLR，2022年。

Allan Zhou、Vikash Kumar、Chelsea Finn 和 Aravind Rajeswaran。面向控制中组合泛化的策略架构。载于*NeurIPS 2022* 深度强化学习研讨会，2022年。

Gaoyue Zhou、Hengkai Pan、Yann LeCun 和 Lerrel Pinto。DINO-WM：基于预训练视觉特征的世界模型支持零样本规划。载于第四十二届国际机器学习会议，2025年。

附录

大语言模型（LLM）的使用

本文部分内容仅使用大语言模型润色文字。

A 扩展背景与相关工作

A.1 多实体环境

更形式化地说，我们将多实体环境 定义为一种可由因子化状态空间表征的环境：$\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2 \otimes ... \otimes \mathcal{S}_N$，其中某些属性在构成实体的不同因子之间共享 $\{S_i\}_{i=1}^N$ 。一个例子是机器人多物体操纵，其中机器人和物体都是实体，每个实体都由位置和视觉外观等共享属性描述（也可能还有其他非共享属性）。另一个例子是机器人运动，其中每个机器人关节都可以视为一个由共享物理属性描述的独立实体。Mohan 等（2024）对（深度）强化学习中的结构进行了全面综述，其中定义并讨论了状态因子化的不同形式。

A.2 面向序贯决策的以物体为中心的表示

许多既有研究探讨了如何利用包含多个实体的环境中的因子化结构进行序贯决策。一些研究假设可以访问系统底层的因子化状态，并开发基于集合的架构，以样本高效且具有良好泛化能力的方式处理这些状态，涵盖无模型（Li 等人，2020；Mambelli 等人，2022；Zadaianchuk 等人，2022；Zhou等人，2022）和基于模型（Sanchez-Gonzalez 等人，2018；Sancaktar 等人，2022）两种设置。为将这些方法扩展到图像观测，一些方法采用端到端学习方式，在学习决策模块的同时学习以物体为中心的表示（Zambaldi 等人，2019；Watters 等人，2019；Veerapaneni 等人，2020；Ferraro 等人，2025）。另一些方法则利用无监督的以物体为中心的表示（OCR），包括基于图像块（Jiang* 等人，2020）、基于槽位（Locatello 等人，2020）和基于粒子（Daniel & Tamar，2022）的模型。这些模型通常会先进行预训练，并在决策组件的下游训练过程中保持冻结。它们已应用于无模型（Zadaianchuk 等人，2021；Yoon 等人，2023；Haramati 等，2024）、基于模型（Zhao 等人，2022；Chang 等人，2023；Mosbach 等人，2024；Zhang 等人，2025）以及模仿学习（Qi 等，2025）方法。

A.3 深度潜在粒子

本节概述了 Deep Latent Particles（DLP，Daniel & Tamar（2022年；2024年））模型；在我们的实验中，我们采用该模型作为无监督以物体为中心的图像表示。DLP 是一种变分自编码器（Variational Auto-Encoder，VAE，Kingma & Welling（2013）），具有由一组称为 潜在粒子 的潜在向量构成的结构化潜在空间。粒子编码图像中的局部区域，这些区域理想情况下对应于物体或物体部件等显著因素。每个粒子表示由以下属性组成：像素空间二维位置 $z_p \in \mathbb{R}^2$、其所表示区域的边界框尺度 $z_s \in \mathbb{R}^2$、用于建模粒子之间遮挡关系的“深度”属性 $z_d \in \mathbb{R}$、透明度 $z_t \in \mathbb{R}$，以及编码粒子区域外观的视觉潜在特征 $z_v \in \mathbb{R}^n$；n 表示每粒子视觉潜在维度超参数。另分配一个粒子用于编码图像背景。粒子数量 M是一个超参数——此前的工作以及本文均据此选择该参数——用于限定图像中感兴趣实体的数量上限。值得注意的是，DLP 完全无监督，因此无法保证每个粒子都表示图像中的一个物体，也无法保证一个物体仅由单个粒子表示。分配大量粒子会提高表示捕获所有感兴趣实体的可能性，但代价是增加 DLP 训练和下游决策过程中的维度与计算复杂度。有关我们实验所用环境中图像的 DLP分解，请参见图 9、10 和 11。

B 环境、任务与数据集

出于透明性和可复现性的考虑，我们详细介绍了本文所使用的环境、任务和数据集。为方便读者，我们着重说明了它们与所依据环境之间的差异。我们介绍了 OGBench 基准（Park 等人，2025年a）中任务的新变体，这些变体使用 MuJoco（Todorov 等人，2012）模拟器实现；同时还介绍了 Chi 等人（2023）提出的 *Push-T* 环境的新变体。OGBench 的实现基于以下官方代码：https://github.com/seohongpark/ogbench）。我们的*PushT* 实现基于 Zhou 等人（2025）提出的一个变体，使用论文提供的官方代码：https://github.com/gaoyuezhou/dino_wm。我们希望，这些突出了组合状态空间和组合泛化等在流行基准中研究较少的挑战的新变体，能够推动实体中心决策及相关领域的进一步研究。数据集和代码（包括环境实现与数据收集脚本）可在 https://sites.google.com/view/hecrl 获取。有关数据收集策略的视频演示，请参见我们的网站。

PPP-Cube***E******nvi***环境：基于 OGBench *Cube* 环境，智能体需要在三维空间中，将立方体从随机初始化的状态操控至目标位置。在每个回合开始时，根据固定的 6 种颜色集合，为每个立方体随机分配不同的颜色。修改后的多色设置有效扩大了状态空间的规模，并且能够在无需泛化到新颜色的情况下，测试随着物体数量增加而实现的泛化能力，以及由此带来的更长时域。图像观测包含两个视图：正面和侧面。为了将算法层面的因素与任何特定的以物体为中心的图像表示相隔离，我们支持基于状态的实体中心观测，以模拟“完美”的因子化表示。具体而言，每个实体（即机器人夹爪和立方体）的状态由其三维位置和偏航角表示。机械臂的状态包括夹爪开合度，而立方体状态则以 0 填充，使其与机器人状态的维度相匹配。每个实体的状态都会与一个独热标识符 $\mathbf{e}_i \in \{0,1\}^7$拼接。$\mathbf{e}_0$ 表示智能体，其余标识符分别表示不同颜色的立方体。像素观测是维度为 128×128 的多视角 RGB 图像。动作空间的维度为 5，包括夹爪三维位置、偏航和开合度的增量。当所有立方体均处于距指定位置的阈值距离以内时，即视为达到目标。

数据集：数据集包含多种智能体与物体的交互，包括通过带噪声的脚本策略采集的**抓**取、**放**置和**推**操作。策略会随机选择一个立方体，执行以下操作之一：(1) 抓取并放置到随机位置，或 (2) 从随机面推动。数据集包含3个立方体，共有300万次状态转移，包含7500个回合，每个回合有400次状态转移。

评估：3个立方体的最大评估回合长度为1000个时间步。我们使用经过250万梯度步训练的检查点，在100个随机采样的初始状态和目标配置上评估每个智能体。

与原版的差异：(1) 像素观测为维度为128×128 的多视角 RGB 图像，而 OGBench 使用维度为 64×64的单视角 RGB 图像。(2) OGBench 将机械臂的视觉属性设为半透明，并用紫色突出显示末端执行器；我们的版本则移除了这一设置。(3) 与 OGBench 相比，状态观测更精简，不包含机器人关节位置和速度、夹爪接触信息以及物体四元数。(4) 在我们的环境中，每个立方体可以呈现6种颜色之一（每次实例化时颜色不重复），无论立方体数量多少；而 OGBench 中的颜色是固定的，且取决于环境中的立方体数量。(5) 我们设置中的数据采集策略有所不同，主要区别在于推操作。(6) 我们数据集中的回合长度为400，而 OGBench 中为1000。

Stack-Cube
环境：参见`PPP-Cube`。
数据集：该数据集包含抓取与放置操作，其中以较高概率进行堆叠；这些操作由带噪声的脚本策略收集。若脚本策略在堆叠方块的过程中移动了堆叠中任意一个较低位置的方块，就会启动恢复机制：将这些方块放回原来的位置，然后再尝试把原先要堆叠的方块放到它们上方。这使数据集更具目标导向性。我们发现，这对于学习能够在3立方体堆叠任务中取得非平凡性能的稳健强化学习策略至关重要（尤其是在我们的评估设置中，达到目标后不会终止回合）。据我们所知，自 OGBench 基准发布以来，还没有方法能在 OGBench 数据集上实现3立方体堆叠任务的非平凡性能

。该数据集包含3个立方体，总计300万次状态转移，分为3000个回合，每个回合包含1000次状态转移。

评估：3个立方体的最大评估回合长度为 1000 个时间步。我们使用经过 300 万梯度步训练的检查点，在 200 个随机采样的初始状态和目标配置上评估每个智能体。

Dif与原文的差异：参见 PPP-Cube 中的 (1)–(4)。(5) 数据采集策略在 ou r setti ng 设置与上文所述不同。

场景

环境：我们采用 OGBench视觉场景环境，其中涉及在三维空间中操控多种物体类型（按钮、立方体、抽屉、窗户）。我们对部分测试任务进行了修改，以避免在更贴近现实的视觉设置中出现目标描述歧义，因为在该设置下机械臂是不透明的。具体而言，我们修改了任务 4 和 5，这些任务的目标涉及立方体位于关闭的抽屉内的图像。在这些情况下，智能体无法推断立方体的目标位置，因为它可能被遮挡。因此，目标改为在抽屉打开并锁定时将立方体放入抽屉。像素观测为单视角 RGB 图像，尺寸为 128×128。动作空间维度为 5，包括夹爪三维位置、偏航和开合的增量。当所有物体均位于距指定位置不超过阈值距离的范围内时，即认为目标已达成。表 5 详细说明了每项任务的内容，其中用 $\rightarrow$ 表示子任务顺序依赖，用 |表示非顺序子任务。

数据集：数据收集策略与 OGBench 相同。带噪声的脚本策略会根据当前状态下各子任务是否适用，按随机顺序执行子任务（例如，只有在抽屉打开时，才会选择将立方体放入抽屉）。数据集共包含 100 万条状态转移，包含 1000 个回合，每个回合有 1000 条状态转移。
评估：评估回合的最大长度为 1000 个时间步。我们使用训练了 1.5M 个梯度步的检查点，对每个智能体进行评估；每项任务使用 50 组经过随机扰动的初始状态和目标配置。

与原版的差异：参见 PPP-Cube 中的（1）–（2）。（3）出于上述原因，我们对任务 4 和 5 做了少许修改。

Push-Tetris***E***

nvi环境：我们将 Chi 等人（2023）提出的*Push-T* 环境改造为用于俄罗斯方块状方块多物体操作的环境。每个回合开始时，系统会随机抽取不同的方块类型。每种方块类型都有一种在所有回合中固定不变的独特颜色。智能体需要推动方块，使其在二维空间中的位置和方向符合目标配置。像素观测为尺寸为 128×128 的单视角 RGB 图像。动作空间维度为 2，包含智能体二维位置的增量。当所有物体的状态-目标像素覆盖率均超过 85% 时，即视为达成目标。达到这一覆盖率需要非常精细的操作能力。我们发现，实践中智能体能够从稀疏奖励中学习，但在测试时成功率较低，尽管终止状态在视觉上与目标相似。因此，我们将覆盖率而非成功率作为主要指标，因为它能提供更多信息。3 个物体时，评估回合的最大长度为 1000 个时间步。

Dataset：数据集通过一种随机策略采集，该策略将活动限制在物体周围的固定半径内，并以 t, sam固定时间间隔进行采样，导致行为高度次优。
评估：3 个物体的最大评估回合长度为 1000 个时间步。我们使用一个经过 100 万个梯度步训练的检查点，在 100 个随机采样的初始状态和目标配置上评估每个智能体。

与原版的区别：（1）原版的 T 形积木被替换为 7 种不同的俄罗斯方块状方块。（2）与原版数据集包含专家示范不同，我们的数据集包含与物体的半随机交互。（3）在我们的设置中，目标由一张独立图像指定；而在原任务中，目标是在与状态观测相同的图像中的阴影区域。

表 5：**详细**场景**任务描述**。我们用 $\rightarrow$ 表示子任务顺序依赖，用 |表示非顺序子任务。

Task	子任务	子任务总数	最长依赖链
1	打开抽屉 | 打开窗户	2	1
2	解锁抽屉 → 关闭抽屉 → 锁上抽屉 | 解锁窗户 → 关闭窗户 → 锁上窗户	6	3
3	打开抽屉 | 解锁窗户 → 关闭窗户 |将立方体移到一侧	4	2
4	打开抽屉 → 锁上抽屉 & 将立方体放入抽屉	3	2
5	解锁抽屉 → 打开抽屉 → 锁上抽屉 & 将立方体放入抽屉 |解锁窗户 → 打开窗户 → 锁上窗户	7	3

B.1 环境难度

我们将本工作考虑的领域与同期研究进行比较，以突出这些领域的复杂性以及我们的方法解决这些领域的有效性。同期研究在 OGBench（Park 等人，2025a）*Cube* 领域的相似或更简单变体上，需要多出几个数量级的数据。Li 等人（2025）在 4 个立方体的环境中需要 $100M$ 次状态转移，才能在单任务强化学习（即非目标条件）设置下完成学习；他们从状态和半稀疏奖励（每个物体成功时给予奖励）中学习，并允许额外的在线交互。我们展示了非平凡的零样本泛化性能：在完全离线、目标条件、稀疏奖励的设置下操作 4 个立方体，同时仅从包含 3 个立方体的 $3M$ 次状态转移中学习，数据既包含状态，也包含图像。Park 等人（2025b）表明，在类似设置下，包括层次化方法在内的非结构化强化学习方法，需要多达 $1B$ 次状态转移，才能在物体数量超过 2 个时取得非平凡性能。据我们所知，根据本文撰写时对所有引用 OGBench 基准的论文所做的审查，迄今尚无其他方法在*Cube* 领域中物体数量超过 2 个时取得非平凡性能。

C 方法实现细节

本节详细介绍我们的方法、基线方法以及实验中使用的图像表示的实现细节。表 9 从多个算法方面对各方法进行了比较。

我们基于官方实现和 LeanRL（https: //github.com/meta-pytorch/LeanRL）使用 PyTorch 实现了所有强化学习方法。LeanRL 是 CleanRL（Huang 等人，2022年）的一个更高效版本。强化学习智能体（即除 EC-Diffuser 以外的所有方法）基于目标条件化的 IQL 变体（Kostrikov 等人，2022年）。低层策略通过 DDPG+BC 从 IQL 的 Q 函数中提取（Fujimoto 和 Gu，2021年）。价值函数训练目标以 0。8 的概率从同一轨迹的未来状态中均匀采样，否则取下一状态。低层策略训练目标从同一轨迹的未来状态中均匀采样（HIQL 除外）。我们在表 6 中报告共享超参数，并在表 7 中报告环境和方法特定的 DDPG+BC 策略提取系数 α 。

表 6：**共享超参数**。

超参数	值
批量大小	512
学习率	0.0003
梯度裁剪范数	20
折扣因子 γ	0.99
目标平滑系数 τ	0.005
IQL/HIQL 期望分位数	0.9
AWR 温度 β	3.0
子目标扩散器扩散步数	10
子目标 K	50
EIT注意力维度	64
EIT注意力头	8
EIT隐藏维度	256
多层感知机层	4
多层感知机隐藏维度	512

EC-SGIQL（我们的方法）：我们基于使用 IQL 训练的 ECRL 智能体（Haramati 等，2024）实现了我们的方法（详情见下文的 EC-IQL）。对于子目标扩散器，我们改造了 EC-Diffuser（Qi 等，2025），使其能够生成以实体中心化状态和目标为条件的子目标。具体而言，我们完全移除动作输入，并将初始状态和目标实体作为干净输入进行条件化，这些输入不会被去噪。我们使用学习得到的加性嵌入，对每个实体分别编码其所属的 3 个集合之一（即状态、目标或带噪子目标）以及多视图设置中的 2 个视图之一。扩散器架构是一个 8 层 Transformer，通过自适应层归一化（AdaLN）以扩散时间步为条件。环境专用扩散器

表 7： **DDPG+BC 策略提取系数** α。

环境	EC-SGIQL（我们的方法）	我们的方法（使用 AWR）	EC-IQL	HIQL	IQL
PPP-Cube (状态)	0.1	0.1	0.1	0.1	0.1
PPP-Cube (图像)	0.2	0.2	0.2	0.3	0.2
Stack-Cube (状态)	0.05	0.05	0.05	0.2	0.05
场景 (图像)	0.3	0.3	0.3	0.4	0.2
Push-Tetris (图像)	0.1	0.1	0.1	0.1	0.1

测试时间超参数详见表8。对于我们的实现，我们改编了 th 中的代码 official EC-Diffuser 仓库：https://github.com/carl-qi/EC-Diffuser。

表8： **子目标扩散器测试时超参数**。

环境	扩散样本 N	子目标步数 T_{sg}	能力半径 $\hat{R}$
PPP-Cube (状态)	256	25	-30
PPP-Cube (图像)	256	25	-30
Stack-Cube (状态)	256	25	-30
场景 (图像)	64	50	-25
Push-Tetris (图像)	64	25	-20

EC-SGIQL with AWR：该方法是我们方法的一项消融实验，因此除子目标策略外，与我们的方法完全相同。这里的子目标策略是一个基于实体交互Transformer（EIT）（Haramati 等，2024）的 Transformer，它用另一个自注意力方块替代了最终的聚合注意力。该方法采用 IQL 价值函数进行训练，优化优势加权回归目标。由于从 DLP粒子中学习时，输出实体和目标实体是无序的，因此我们使用倒角距离作为损失，而非标准均方误差。

EC-IQL：对应于一个 ECRL（Haramati 等，2024）智能体，该智能体通过将实体交互 Transformer（EIT）架构与 IQL 相结合，适配到离线设置中。我们对用于 Q函数的 EIT 架构稍作修改，以 AdaLN 条件化替代动作实体条件化。我们的实现基于官方 ECRL代码库：https://github.com/ DanHrmti/ECRL。

HIQL：此基线对应于一个基于 HIQL（Park 等人，2023）的智能体。我们调整了官方实现，使其与其他方法相似，以便进行公平比较，同时保留核心属性不变。具体而言，我们使用底层 IQL 智能体进行训练，并通过 DDPG+BC 提取低层策略。与本文中的其他基于 IQL 的方法不同，与原始 HIQL 一样，我们在子目标分布上而非目标分布上训练低层策略。高层策略使用 IQL价值函数，通过 AWR 进行训练。在测试时，子目标保持的时长与 Diffuser子目标相同（见表8）。所有智能体组件均为隐藏维度为 512 的 4 层多层感知机。我们的实现基于 OGBench 官方实现：https://github.com/seohongpark/ogbench。

IQL：该基线对应一个基于 IQL（Kostrikov 等人，2022年）的目标条件智能体。智能体的所有组件均为 4 层多层感知机，隐藏维度为 512。我们的实现基于 OGBench 的官方实现：https://github.com/seohongpark/ ogbench。

EC-Diffuser（Qi 等，2025）：是一种以实体为中心的扩散式行为克隆方法。它采用基于 Transformer 的架构，以当前状态和目标状态作为输入，并通过扩散预测未来状态和动作。在控制方面，它联合预测未来状态和动作，并以模型预测控制的方式执行第一个去噪后的动作。对于我们的所有任务，我们使用的预测时域为 5，扩散步数为 10。对于所有其他超参数，我们使用所报告的值。我们使用的官方实现来自：https://github.com/carl-qi/EC-Diffuser。

表 9：**不同方法在算法方面的比较**。

属性	EC-SGIQL（本文方法）	本文方法 w. AWR	EC-IQL	EC-Diffuser	HIQL	IQL
RL	✓	✓	✓	✗	✓	✓
实体中心化	✓	✓	✓	✓	✗	✗
层次化	✓	✓	✗	✗	✓	✗
观测扩散	✓	✗	✗	✓	✗	✗

图像表示：我们分别在表 10 和表 11 中给出了 DLPv2（Daniel 与 Tamar，2024）和 VQ-VAE（Van Den Oord 等人，2017年）模型的超参数。我们使用的实现来自：https://github.com/DanHrmti/ECRL。

表10：**DLP超参数**。

环境	先验粒子	后验粒子	视觉特征维度	背景粒子维度
PPP-Cube（图像）	32（每个视角）	20（每个视图）	8	1
场景 (图像)	32	24	8	1
Push-Tetris (图像)	32	20	8	1

表11：**VQ-VAE 超参数**。

超参数	值
嵌入维度	16
字典大小	2048
展平潜在维度	1024

D 补充结果与讨论

D.1 在层级结构中解耦训练目标分布的益处

在我们的目标条件设定中，价值函数可被视为状态之间的一种折扣（非对称）距离指标（Wang 等人，2023）。累积误差与折扣因子共同定义了一个“有效半径”，其中的值能为策略提取提供清晰的学习信号；我们将该半径称为策略能力半径，R（见第 3.1 节）。由于我们无法获得真实的价值近似误差，该量事先未知，并且在不评估已提取策略的情况下，尚不清楚它在实践中能否被恢复或有效近似。以往涵盖离线强化学习（Park 等人，2023）和扩散规划（Li 等人，2023；Chen 等人，2024）的层次化方法，会固定单一超参数 K，用以定义层级结构两个层级的训练分布：低层策略在相距至多 K 个时间步的状态-目标对上训练，而高层策略则被训练为生成这些 K 步状态作为子目标。选择 K 体现了复杂度与样本效率之间的权衡：较小的 K 会简化策略目标分布，但通过限制策略所训练的状态-目标对，对离线数据的利用更少，从而可能阻碍泛化。因此，选择 K 需要在层级结构的所有层级上同时平衡这一权衡。

我们选择了一种替代方案：在完整的目标分布上训练低层策略，也就是价值函数所使用的相同分布。此选择将层级之间对训练目标的依赖解耦，使我们能够通过将超参数 R 调整为测试时约束（详见第 3.2 节），隐式地为低层策略选择 K^；相比之下，调整 K 并不理想，因为这无需重新训练。随后，可以更灵活地为子目标生成器选择 K，以此为低层策略的能力设定上限 R，同时权衡复杂度。

D.2 长时域物体操作

中间成功情况（例如，已完成的子任务数量）见表12，Scene任务的详细结果见表13。在 Scene 的大多数任务上，EC-IQL 的表现与我们的方法相当，总体性能排名第二。我们认为，这是因为以实体为中心的结构有助于学习准确且在内部进行因子分解的价值函数。Scene 的组合复杂度最低，这或许可以解释为什么层级结构带来的益处较小：因子化价值函数在这一领域提供了良好的学习信号。EC-Diffuser 在 Stack-Cube 上的表现与我们的方法相当。我们认为，这主要是由于该数据集的特性：与其他数据集相比，它更以目标为导向；此外，它不依赖价值函数，因此不会面临与强化学习方法相同的长时程挑战。HIQL 和 IQL 在利用单向量表示学习时，难以扩展到包含多个物体的基于图像的任务；这与以物体为中心的决策文献中的既有发现，以及 OGBench 基准上的结果一致（Park 等人，2025a）。在某些情况下，HIQL 的表现甚至不如 IQL；我们认为，这是因为使用 AWR 训练子目标策略往往会产生质量较低的子目标（见第 4.4 节和 D.5）。

表12：**长时程操作性能**。除 Push-Tetris 外，所有值均为成功率；对于 Push-Tetris，我们报告状态目标像素倒角距离。最佳结果（误差范围为一个标准差）以 **粗体**标出。

环境	EC-SGIQL	EC-IQL	EC-Diffuser	HIQL	IQL
PPP-Cube （状态）	**88.0** ± 1.4	60.8 ± 4.5	66.5 ± 7.0	62.8 ± 6.5	46.5 ± 7.0
PPP-Cube (图像)	**82.3** ± 2.2	60.8 ± 1.5	9.5 ± 1.3	0.0 ± 0.0	0.0 ± 0.0
Stack-Cube （状态）	**58.8** ± 1.7	43.0 ± 4.6	**55.3** ± 6.9	6.0 ± 0.8	38.5 ± 1.0
场景 (图像)	**90.0** ± 2.2	84.0 ± 2.7	51.5 ± 0.6	65.3 ± 1.7	77.3 ± 0.5
Push-Tetris (图像)	**19.0** ± 3.5	54.7 ± 4.9	84.2 ± 3.6	94。1 ± 0.9	98.8 ± 1.8

表13：**详细成功率（针对**场景**任务**）。在一个标准差范围内的最佳结果以**粗体**突出显示。

Task	EC-SGIQL	EC-IQL	EC-Diffuser	HIQL	IQL
1	**85.0** ± 6.0	**83.5** ± 4.4	14.0 ± 14.3	20.0 ± 5.2	27。0 ± 4.8
2	83.5 ± 1.9	**89.0** ± 2.0	6.0 ± 3.3	15.0 ± 4.8	30.5 ± 16.7
3	**52**。**0** ± 7.1	16.5 ± 13.4	0.0 ± 0.0	7.0 ± 2.6	15.0 ± 4.8
4	**43**。**0** ± 10.5	**39.0** ± 15.9	0.5 ± 1.0	0。5 ± 1.0	6.0 ± 1.6
5	**44.0** ± 17.2	**37.0** ± 7.8	0.0 ± 0.0	0.0 ± 0.0	10.0 ± 7.1

D.3 泛化

本文所比较的以实体为中心的方法的组合泛化结果见表14。在包含4、5、6和7个物体的泛化实验中，最大评估回合长度分别为1200、1500、2000和2000。我们还将推理粒子数量——DLP支持的一项特性——分别增加到24、28、30和30。

D.4 子目标扩散器超参数敏感性研究

我们在本套件中最具挑战性的任务PPP-Cube *(Image)*上，测试了本方法对子目标相关超参数 K、T_{sg} 和 N 的敏感性。结果见图5。我们发现，该方法对这些超参数大体上具有稳健性，在所有测试配置下均显著优于底层强化学习智能体。

K （图5，左）：结果与我们的直觉一致，即按可达性筛选子目标可使本方法对 K 保持稳健，只要它是策略能力半径 R 的上界（见附录D.1中的讨论）。当 $K = 10$时，性能下降，这表明能力半径大于该值。

T_{sg} (图5，中)：较大的值会缩短平均动作执行时间，因为所需的

表14：**零样本组合泛化**。除 Push-Tetris 变体报告状态-目标像素覆盖率外，所有值均为成功率。* 表示每个分段中的训练变体，该变体用于评估其他变体上的零样本能力。与最佳结果相差不超过一个标准差的结果以**粗体**突出显示。

变体	EC-SGIQL	EC-IQL	EC-Diffuser
2-Cubes-Stack (状态)	69.8 ± 7.0	45.8 ± 12.3	**87.8 ± 5.4**
3-Cubes-Stack* (状态)	**43.5 ± 1.9**	29.0 ± 2.9	**43。8 ± 9.2**
4立方体堆叠-2 (状态)	**28.8 ± 3.9**	17.3 ± 2.5	1.5 ± 1.3
PPP-1立方体 (状态)	**97.0 ± 1.8**	**92.3 ± 8.9**	85.0 ± 9.2
PPP-2立方体 (状态)	**94.0 ± 3.4**	75。0 ± 3.8	76.5 ± 7.6
PPP-3-立方体* (状态)	**82。5 ± 3.1**	51.5 ± 4.4	44.8 ± 6.7
PPP-4-立方体 (状态)	**65.3 ± 3.5**	31.8 ± 4.0	42。8 ± 11.4
PPP-5-立方体 (状态)	**49.0 ± 8.5**	19.3 ± 5.9	23.8 ± 6.7
PPP-6-立方体 (状态)	**25.7 ± 1.5**	10.5 ± 6.1	8。8 ± 3.5
PPP-1立方体 (图像)	**94.5 ± 2.9**	**91.5 ± 1.3**	3.8 ± 1.3
PPP-2立方体 (图像)	**77。0 ± 3.6**	47.8 ± 6.2	0.0 ± 0.0
PPP-3-立方体* (图像)	**64.3 ± 6.0**	25.0 ± 5.7	0.3 ± 0.5
PPP-4-立方体 (图像)	**38.3 ± 5.7**	11.5 ± 4.7	0.0 ± 0.0
PPP-5-立方体 (图像)	**19.3 ± 6.2**	4.0 ± 2.0	0.0 ± 0.0
PPP-6-立方体 (图像)	**9.5 ± 1.3**	1.5 ± 1.3	0.0 ± 0.0
Push-Tetris-单物体 (图像)	**8。0 ± 1.5**	**6.6 ± 2.6**	5.0 ± 1.8
Push-Tetris-双物体 (图像)	**61.4 ± 5.7**	40.8 ± 4.2	6.3 ± 0.4
Push-Tetris-三物体* (图像)	**61.4 ± 3.3**	31.6 ± 1.3	7.9 ± 0.5
Push-Tetris-4-Objects (图像)	**52.4 ± 3.5**	20.7 ± 1.7	6.4 ± 1.1
Push-Tetris-5-Objects (图像)	**41.8 ± 6.0**	14.4 ± 2.3	4.6 ± 0.6
Push-Tetris-6-Objects (图像)	**32.0 ± 1.8**	11.5 ± 2.2	4.6 ± 1.2
Push-Tetris-7-Objects (图像)	**21.1 ± 2.2**	10.0 ± 1.5	4。7 ± 0.8

扩散生成次数更少；但如果该值远大于强化学习策略到达子目标所需的值，决策时间步效率可能会降低。此外，鉴于采样得到的子目标并不完美，以合理的频率重新生成子目标对智能体有所助益。这类似于模型预测控制中的重新规划。我们发现，值较大时，性能开始下降。

N（图5，右）：较小的样本量会降低采样到“最优”子目标的概率，从而导致性能方差较高。较大的样本量会增加生成被错误赋予高价值的子目标的概率，从而可能降低性能。介于 $64 - 256$ 之间的中等样本量可带来同样高的成功率。

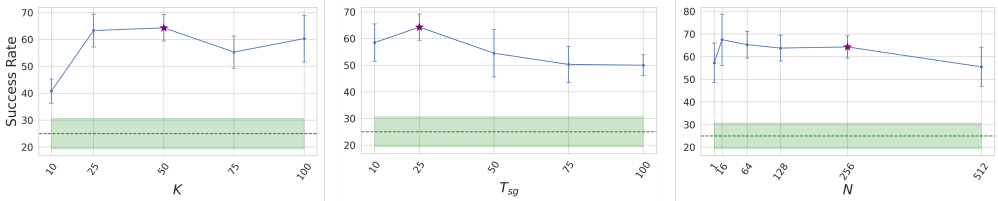

图5：**超参数敏感性研究**。我们的方法在 PPP-Cube 上的成功率(图像)。每幅图中只改变一个超参数，其余参数与主要结果所用的参数保持一致，我们将这些参数称为默认值。紫色星号表示默认值，误差线表示标准差，绿色虚线和阴影区域表示无子目标 *EC-IQL* 基线的均值和标准差。**左图**：改变 K，即子目标扩散器训练分布中状态与子目标之间的时间步差。**中图**：改变 T_{sg}，即测试时子目标展开的时间步数。**右图**：改变 N，即筛选前采样的测试时扩散子目标数量。

D.5 AWR 子目标可视化

我们在图6、图7和图8中提供了在 DLP 表征上使用 AWR 训练的确定性实体中心 Transformer 所生成子目标的可视化。我们观察到，在所有环境中，子目标常常包含实体的重复，从而为低层 GCRL 策略提供了含糊的子目标。这种重复可以用 DLP 表征中的冗余以及优势加权回归目标来解释。如第4.5节所述，加权回归拟合的是未来观测的加权平均；而在我们的“play”数据集中，给定初始状态和目标，存在许多可能的未来。DLP 用许多粒子表示一个场景；由于其无监督特性，可能用多个粒子表示同一物体（见附录 A.3）。这使得每个粒子都能在同一物体 的单个子目标中捕捉不同的未来子目标，而这正是我们在这些图中观察到的现象。子目标平均现象并非纯粹由 DLP 表征的冗余造成，因为它在基于状态的环境中同样出现（见第4.5节，表3）。这些结果凸显了数据中的多模态性，并有助于解释使用 AWR 训练子目标生成器的基线方法和消融实验性能较低的原因。这种现象在我们的子目标扩散器中很少发生（即便发生，严重程度也低得多）；子目标扩散器能够捕捉数据中的多种模式，并支持在测试时分别从这些模式中采样。我们认为，更大的扩散 Transformer 可能有助于完全缓解这一现象。

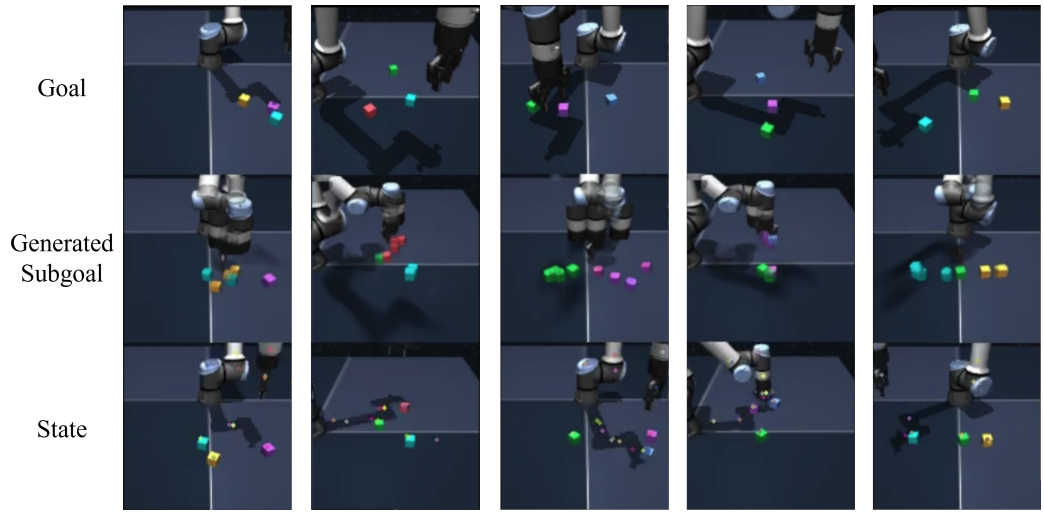

图 6：PPP-Cube**AWR 子目标**。子目标（中间）由 DLP 解码器重建，并根据状态的 DLP 表征（底部）和目标（顶部）生成。各列并非按顺序排列，即每列表示彼此无关的子目标。

D.6 预训练图像表征重建

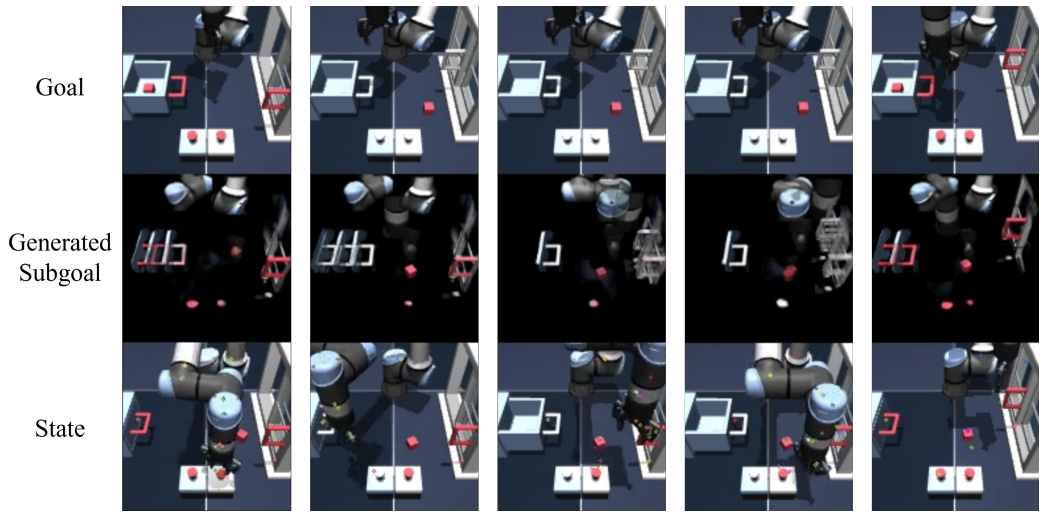

图 7：场景 **AWR 子目标**。子目标（中部）由 DLP 解码器重建，并根据状态（底部）和目标（顶部）的 DLP 表征生成。各列并非按顺序排列，即每一列代表彼此无关的子目标。

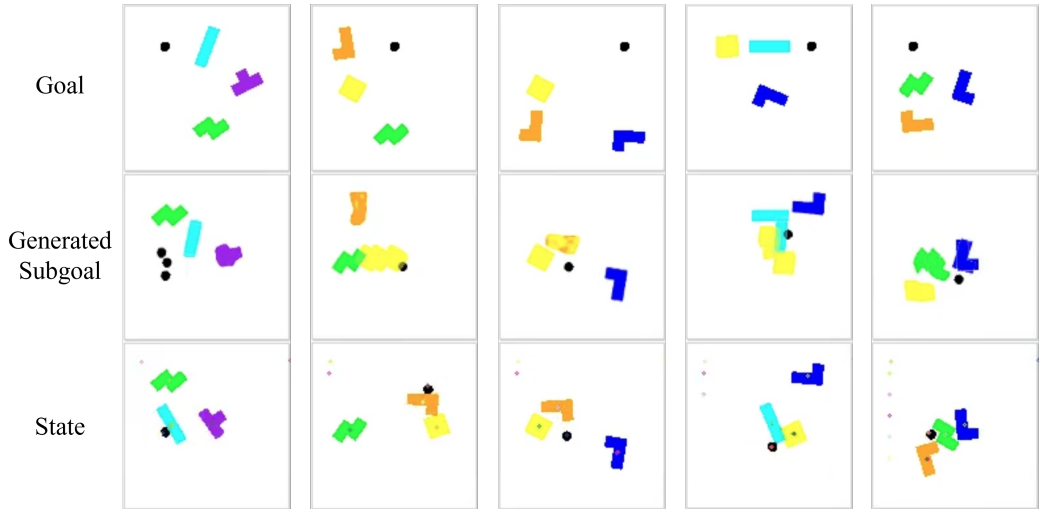

图 8：Push-Tetris **AWR 子目标**。子目标（中部）由 DLP 解码器重建，并根据状态（底部）和目标（顶部）的 DLP 表征生成。各列并非按顺序排列，即每一列代表彼此无关的子目标。

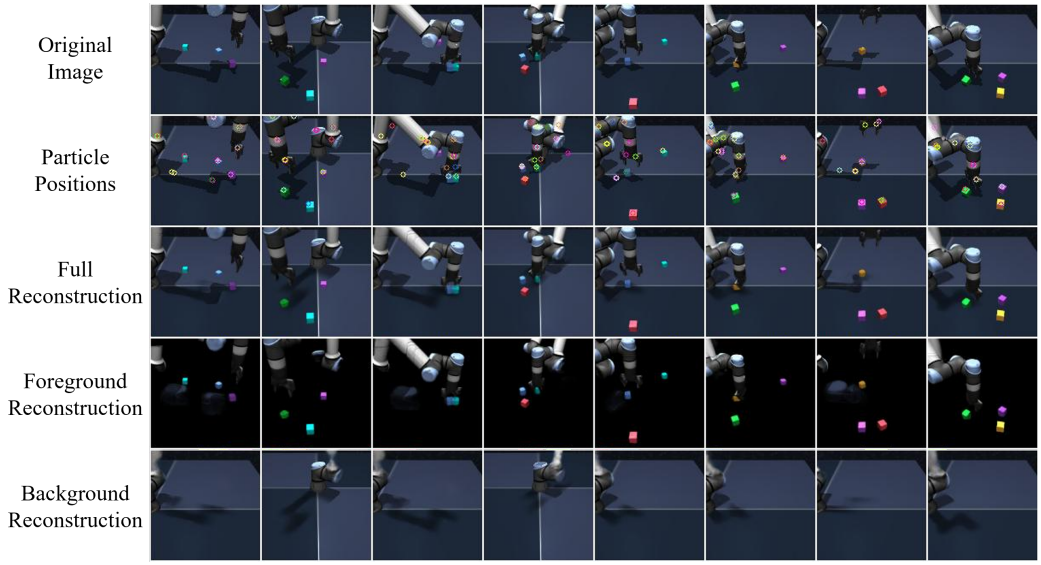

图9： `PPP-Cube` **DLP 分解**。

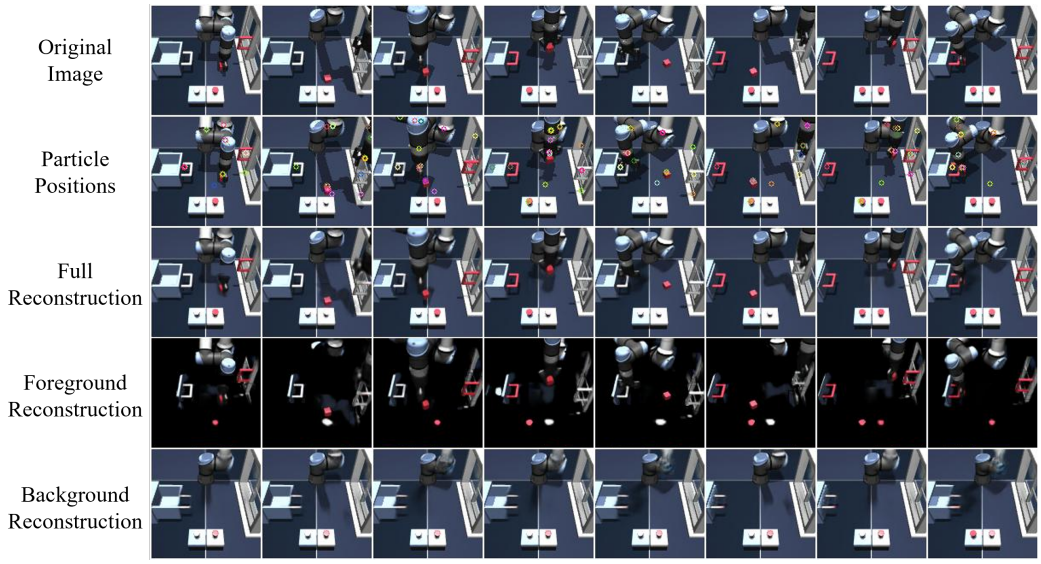

图10： 场景**DLP 分解**。

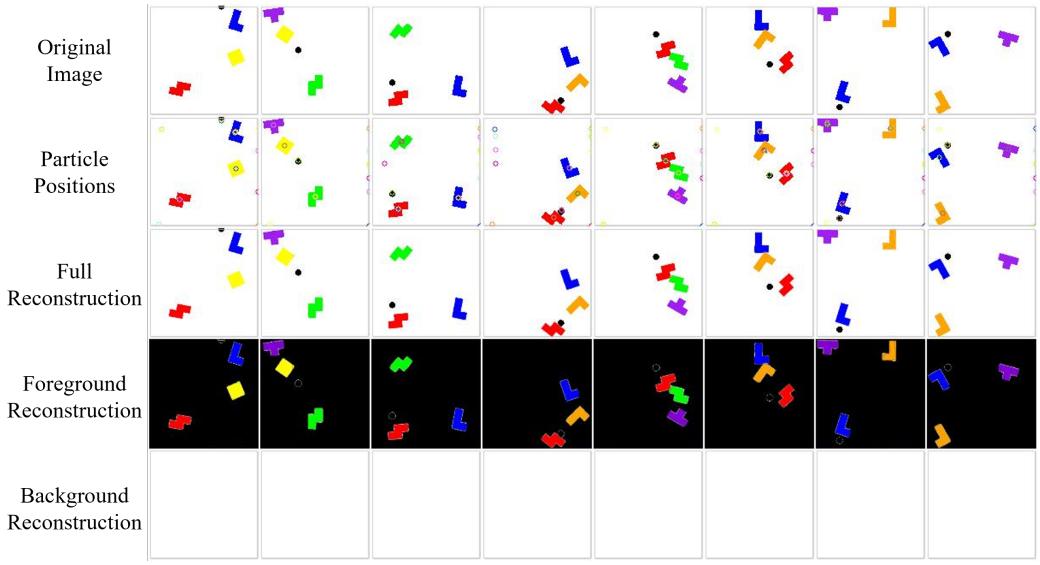

图11： `Push-Tetris` **DLP 分解**。

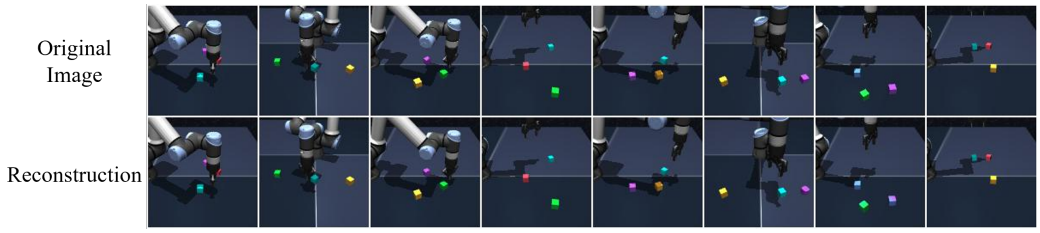

图12： `PPP-Cube` **VQ-VAE 重建**。

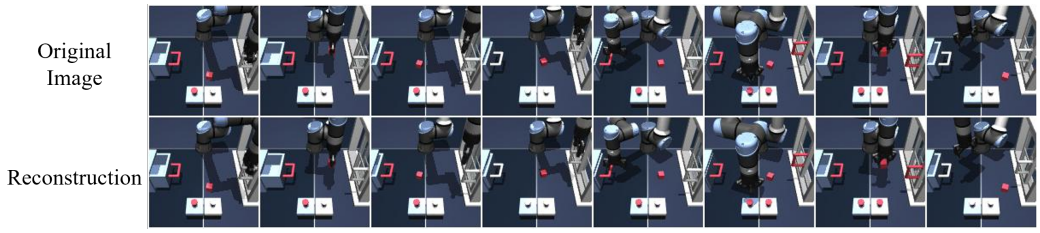

图13： 场景**VQ-VAE 重建**。

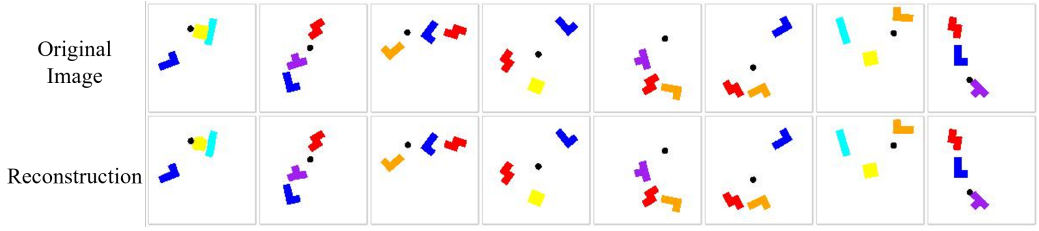

图14： `Push-Tetris` **VQ-VAE 重建**。