arXiv:2611.32776v1 [cs.AI] 24 Nov 2026

通过动作的对比识别学习世界模型

Hayeong Lee 高丽大学 首尔，大韩民国
hayeong lee@korea.ac.kr

Youngjun Suh 高丽大学 首尔，大韩民国
youngjunsuh@korea.ac.kr

Byung-Jun Lee 高丽大学，Gauss Labs Inc.，首尔，大韩民国byungjunlee@korea.ac.kr

摘要

潜在世界模型支持基于离线视觉数据进行目标导向规划，但仅靠预测并不能明确其表征应保留哪些与动作相关的差异。逆动力学提供辅助监督，但平方误差动作回归只会通过条件动作均值的变化来奖励额外信息。我们提出动作对比式识别（*CIA*），通过从历史条件行为提议中采样备选项，识别已执行动作。比较同一情境下可能采取的动作，可以针对下一个潜变量在历史之外所提供的信息。在行为完全匹配的条件下，我们证明，*CIA* 能够奖励那些不会改变最优逆回归误差的动作信息，并通过一个受控学习示例验证了这一区别。在五个视觉控制基准上，*CIA* 的规划器平均成功率高于基础世界模型和回归逆动力学基线，且无需修改规划器。

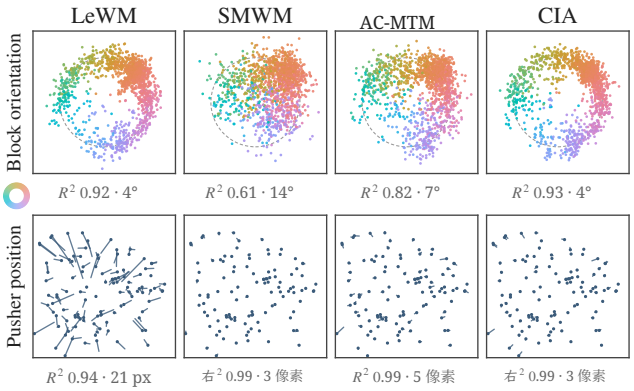


图1。**从冻结的 PushT 表征中解码出的目标组成部分**。上：预测的方块角度坐标；准确的读数按真实角度的颜色顺序沿单位圆分布。下：线段越短，推杆定位越准确。CIA 支持对这两个组成部分进行准确读数。数字表示测试集R^2 和中位误差。进一步分析见第5.4节。

1. 引言

世界模型通过预测动作后果，支持基于无奖励离线数据的目标导向控制 [8,20]。联合嵌入预测架构（JEPAs）在潜在空间中学习这些预测 [1, 14, 27]。随后，规划器会搜索预测潜变量，使其接近目标图像的潜变量。因此，只有当潜变量保留了区分目标与其他可达状态的信息时，规划才能成功。

仅靠潜在预测无法充分刻画这些差异。由于编码器也定义了预测目标，某个观测分量可能很容易预测，却对控制毫无用处。防坍塌正则化能够保留变化，但无法确定哪些变化对动作选择至关重要 [2, 14]。辅助目标则通过潜在轨迹几何结构 [7, 25] 或逆动力学 [3, 10]提供进一步的判据。逆动力学非常适用于离线数据：每个记录的转移都带有其已执行动作这一任务无关标签，它将动作与其效果联系起来。

这种监督应奖励哪些动作信息？平方误差逆回归通过条件动作均值的变化，对下一个潜变量进行奖励。我们表明，即使增加动作信息，最小期望误差也可能保持不变。例如，即使条件动作均值保持不变，下一个潜变量表征仍可能揭示动作的幅度。这促使我们采用能够奖励更广泛动作—结果区分的监督方式。

我们提出<风格 id='1'>动作的对比识别 </风格>（CIA），这是一种用于在世界模型表示中保留动作信息的辅助目标。它从学习到的历史条件行为模型所生成的备选项中，学习识别已执行的动作，从而促使下一个潜变量保留超出历史本身所提供信息的动作信息。在条件 InfoNCE [21], 的基础上，我们证明，当提议分布与离线数据中的历史条件动作分布完全匹配时，只要下一个潜变量提供了超出历史的动作信息，即使条件动作均值保持不变，最优对比损失也会严格低于随机机会水平的损失。

条件动作分布时，只要下一个潜变量提供了超出历史的动作信息，即使条件动作均值保持不变，最优对比损失也会严格低于随机机会水平的损失。

CIA 更新由前向预测和目标表示共享的编码器。我们假设，保留不同合理动作结果之间的区别，有助于规划器将预测轨迹与目标进行比较。该提案和评分仅在训练期间需要；规划时仍保留常规预测器和潜在目标代价。图 1展示了相应的实证优势：在 Push 中，CIA 仍能准确获取推动器位置和方块方向 [4]。

在五个视觉控制基准和两个规划器上，CIA 的总体平均成功率均高于 LeWM 和回归 IDM 基线。这些结果证明了 CIA 作为潜在世界模型辅助目标的实际价值。将 CIA 与 CGS [7] 结合，可进一步提高 PushT 和 Cube 上的平均成功率 [16], ，这表明将对比动作识别与几何正则化结合具有潜力。候选来源消融实验、辅助网络诊断和目标组成部分分析，考察了促成这些提升的设计选择与结果。

我们的贡献主要有三点：
•我们提出 CIA，这是一种融入行为信息的逆向对比目标，用于世界模型学习，且不改变规划流程。

• 我们刻画了平方误差逆向回归的局限性，并展示了如何通过与行为匹配的替代项进行对比动作识别，从而奖励额外的动作信息。

• 我们展示了在五个视觉控制基准上规划效果的提升。对替代项来源的消融实验以及对规划失败的分析，揭示了性能提升的来源。

2. 相关工作

世界模型表示学习。JEPA 风格的世界模型通过潜变量预测支持规划 [1, 27], ，而 LeWorldModel（LeWM）将预测与 SIGReg 结合起来 [14]。TS 和 CGS 分别塑造轨迹曲率和动作—位移几何 [7, 25]。SMWM 对连续潜变量中的动作进行回归，以防止表征坍塌 [10]。与 SMWM 一样，CIA 使用已执行动作训练编码器，但它是在备选项中识别这些动作，而不是对其进行回归。

对比式世界模型。 C-SWM 通过对比式状态分离学习以对象为结构的动力学 [11]，而时间预测编码则将观测到的未来—历史配对与无关轨迹进行对比 [15]。AC-MTM 从批次中其他状态转移的动作里识别出已执行动作 [3]。与这些批次内的备选项不同，CIA 的备选项是针对同一段历史抽取的，因此在该历史下是合理的。

逆向信息与控制。 GLAMOR 在序列上学习逆向模型和动作先验模型，并利用它们的似然比进行规划 [17]。Rakelly 等人 [19] 研究了信息目标能否产生足以实现控制的表征；他们的逆向目标 $I(A_t; Z_{t+k} \mid Z_t)$在 $H = Z_t$ 时包含了我们的单步目标。我们仅使用相关比值（公式（14））来训练潜变量，并探究哪些动作信息对比式识别奖励是回归所无法提供的，而不是探究其是否足以实现控制。

3. 预备知识

我们从一个不包含奖励或任务标签 [1, 14, 25, 27] 的观测-动作轨迹离线数据集中学习编码器和潜在动力学。在测试集时，规划器利用它们，根据给定的图像目标进行到达。

对于长度为 ℓ 的历史，我们构建轨迹窗口离线数据集，

$$\mathcal{D} = \left\{ \left(o_{t-\ell+1:t+1}, a_{t-\ell+1:t} \right) \right\}_{t \in \mathcal{I}}, \tag{1}$$

其中 $x_{i:j} = (x_i, \ldots\ldots, x_j)$，而 $\mathcal{I}$ 对单个回合内的 N 个有效窗口进行索引。观测满足 $o_t \in \mathcal{O}$，连续动作满足 $a_t \in \mathcal{A} \subseteq \mathbb{R}^{d_a}$。这里， a_t 表示与从 o_t 到 o_{t+1} 的状态转移相关联的控制输入。编码器 f_θ： $\mathcal{O} \to \mathcal{Z} \subseteq \mathbb{R}^d$ 生成 $z_t = f_\theta(o_t)$ 。动作前的潜在历史为

$$h_t = \left(z_{t-\ell+1:t}, a_{t-\ell+1:t-1} \right) \in \mathcal{H}, \tag{2}$$

其中不包括 a_t 和 o_{t+1}。对于 $\ell = 1$，过去的动作序列为空，且 $h_t = z_t$。前向模型g_ϕ： $\mathcal{H} \times \mathcal{A} \to \mathcal{Z}$ 预测 $\hat{z}_{t+1} = g_\phi(h_t, a_t)$。

基于 JEPA 的世界模型。遵循 LeWM [14], ，我们通过一步潜在预测联合训练编码器和预测器：

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{\mathcal{D}} \left[\left\| g_\phi(h_t, a_t) - f_\theta(o_{t+1}) \right\|_2^2 \right], \tag{3}$$

其中，期望是对从离线数据集中采样的轨迹窗口取的。

为避免表示坍塌，我们采用草图化等方高斯分布正则化（SIGReg） [2], ，通过随机一维投影促使潜在特征服从等方高斯分布。基础世界模型目标为

$$\mathcal{L}_{\text{WM}} = \mathcal{L}_{\text{pred}} + \alpha \, \mathcal{L}_{\text{SIGReg}}, \tag{4}$$

其中 $\alpha > 0$ 控制正则化强度。该目标并未明确要求在同一历史下可能发生的备选项中识别出已执行动作。

遵循 LeWM [14], , 我们在学习值潜在空间中执行轨迹优化。给定当前历史和候选动作序列 $a_{t:t+T-1}$，预测器 g_ϕ会在规划时域内自回归地滚动预测未来潜变量 T。我们对动作序列进行优化，使终端预测潜变量与目标嵌入之间的欧几里得距离平方最小 $z_g = f_\theta(o_g)$：

$$a^*_{t:t+T-1} = \arg\min_{a_{t:t+T-1}} \|\hat{z}_{t+T} - z_g\|_2^2 . \tag{5}$$

编码器和预测器在规划过程中保持固定。

概率形式化。为便于分析，我们省略时间索引，并将从 $\mathcal{D}$ 中抽取的窗口视为随机变量：H、 A和 $Z' = f_\theta(O')$分别表示其历史、已执行动作和下一个潜变量，小写的 h、 a、 z' 表示它们的取值。对于固定编码器 f_θ，数据诱导出联合分布

$$p(h, a, z') = p(h)\, p_\beta(a \mid h)\, p(z' \mid h, a), \tag{6}$$

其中， $p_\beta(a \mid h)$ 是行为分布，即给定历史时已执行动作的条件分布。项 $p(z' \mid h,\ a)$ 表示下一个潜在表示所对应的条件分布。

我们假设这些分布相对于固定的参考测度都具有密度，并且所有积分均被理解为相对于这些测度而言。下一个潜变量的边缘分布为

$$p(z' \mid h) = \int_{\mathcal{A}} p(z' \mid h, a)\, p_\beta(a \mid h)\, \mathrm{d}a. \tag{7}$$

根据贝叶斯定理，在 $p(z' \mid h) > 0$ 的情况下，动作后验分布为

$$p_\beta(a \mid h, z') = \frac{p(z' \mid h, a)\, p_\beta(a \mid h)}{p(z' \mid h)}. \tag{8}$$

后验结合了状态转移证据和行为先验。我们的分析固定编码器，并始终使用相同的条件变量 H 。

4. 方法

我们提出动作的对比识别 （CIA），这是一种辅助目标，通过学习识别已执行动作来补充潜在预测。我们的设计结合了两种选择：将已执行动作与其他候选项进行对比，而非回归其取值；并从根据历史条件确定的行为分布中拟合出的提议分布中抽取这些候选项。我们首先解释回归所奖励的信息，然后展示这两种选择如何使 CIA 能够奖励更多动作—结果差异。

4.1. 回归逆动力学

回归式 IDM 根据动作执行前的信息和下一个潜变量来预测动作：

$$\mathcal{L}_{\mathrm{IDM}} = \mathbb{E}\|A - r(H, Z')\|_2^2. \tag{9}$$

其中 $r(h, z') \in \mathbb{R}^{d_a}$ 是回归头。为了了解这一损失所鼓励的内容，我们比较在三种输入信息水平下可达到的最低误差：

$$\begin{aligned} V &:= \mathbb{E}\|A - \mathbb{E}[A]\|_2^2, \\ V_H &:= \inf_u \mathbb{E}\|A - u(H)\|_2^2, \\ V_{H,Z'} &:= \inf_r \mathbb{E}\|A - r(H, Z')\|_2^2. \end{aligned} \tag{10}$$

V 是最佳常数猜测 $\mathbb{E}[A]$的误差，因此等于动作总方差。V_H 是仅观察历史的预测器 $u(h) \in \mathbb{R}^{d_a}$ 所能达到的最低误差，而 $V_{H,\ Z'}$ 是同时观察下一个潜变量的预测器 $r(h,\ z')$ 所能达到的最低误差。下确界取遍所有具有有限误差的可测预测器，因此这些是任何预测头都能达到的最佳误差。我们假设 $0 < V < \infty$，并令 $R_H^2 := 1 - V_H/V$表示仅由历史所解释的动作方差占比。V_H 和$V_{H,\ Z'}$ 必须使用相同的历史 H：SMWM 使用 $H = Z_t$ [10]，而更长的历史会定义出不同的比较对象。

命题 1（下一个潜变量的贡献）。观察 Z' 所带来的最小误差降低量为

$$\begin{aligned} \Delta &:= V_H - V_{H,Z'} \\ &= \mathbb{E}\big\|\mathbb{E}[A \mid H, Z'] - \mathbb{E}[A \mid H]\big\|_2^2, \end{aligned} \tag{11}$$

with

$$0 \le \Delta \le V_H = (1 - R_H^2)V. \tag{12}$$

我们将 Δ 称为 均值偏移：即 Z'对条件动作均值造成的预期变化。在平方误差回归中，下一个潜变量只能通过这一均值偏移产生作用。该界限进一步表明，这一作用至多为 $V_H = (1 - R_H^2)\ V$，即历史未能解释的动作方差；当历史已经能够很好地预测动作时，下一个潜变量在回归中的收益就可能很小。该陈述针对的是最优误差，而不是经过训练的预测头是否使用 Z'。附录 A 给出了证明及相关限定条件。

4.2. 条件逆对比目标

我们在相同历史条件下，从备选项中识别已执行动作。从数据中抽取 $(H,\ A_0,\ Z')$，其中 $A_0 = A$；再从 $q(\cdot \mid H)$ 中独立抽取 $K - 1$ 个备选项，且在给定 H 的条件下，这些备选项与 $(A_0,\ Z')$ 相互独立。对于$K \ge 2$，共享的逐候选项评分 S_ψ 最小化

$$\mathcal{L}_{\mathrm{CIA}}(S_\psi; q) := \mathbb{E}\left[-\log \frac{\exp S_\psi(H, Z', A_0)}{\sum_{j=0}^{K-1} \exp S_\psi(H, Z', A_j)}\right]. \tag{13}$$

算法 1 动作的对比识别

要求： 离线轨迹 $\mathcal{D}$、候选数量 K、权重 λ

初始化 f_θ、g_ϕ、S_ψ 和 q_η
预拟合 η 使用（20），并冻结世界模型
对每次训练迭代**执行**
  采样轨迹窗口及其执行的动作 a 自 $\mathcal{D}$
  构造 h 以及观测到的下一潜变量 $z' = f_\theta(o')$
  设置 $a_0 = a$；从 $K-1$ 中抽取并分离 $K-1$ 备选项来自更新前的$q_\eta(\cdot \mid \mathrm{sg}(h))$
  更新 η 仅使用（20）
  联合置换存储的候选项和正标签
  使用相同的13）计算 z' 以处理每个候选项
  使用（更新（θ，ϕ，ψ））21，固定 q和候选项-项不变
结束循环
返回 f_θ，g_ϕ 用于规划

所有候选项使用相同的观测到的下一个潜变量；无需反事实结果。

在附录中的密度和支撑集假设下B，贝叶斯最优逻辑是

$$S^*(h, z', a) = \log \frac{p_\beta(a \mid h, z')}{q(a \mid h)} + c(h, z') = \log p(z' \mid h, a) + \log \frac{p_\beta(a \mid h)}{q(a \mid h)} + \tilde{c}(h, z'), \tag{14}$$

其中，加性项与 a无关。当$q(a \mid h) = p_\beta(a \mid h)$时，行为先验的对数比消失。此时，最优评分根据给定历史和各个候选动作时观测到下一个潜变量的似然 $p(z' \mid h, a)$对候选动作进行排序。这启发我们将历史条件提议分布拟合到离线动作上。对于拟合后的提议分布，同一表达式描述了相对于 q的兼容性，其中包括任何剩余的先验比贡献。此构造建立在对比估计和条件 InfoNCE [13, 21, 23]的基础上。

命题2（超越历史的动作识别）。若$q(\cdot \mid H) = p_\beta(\cdot \mid H)$几乎处处成立，则每个忽略 Z' 的评分 $s(H, A)$ 都满足

$$\mathcal{L}_{\mathrm{CIA}}(s; p_\beta) \geq \log K. \tag{15}$$

对于每个评分 S，

$$I(A; Z' \mid H) \geq \log K - \mathcal{L}_{\mathrm{CIA}}(S; p_\beta). \tag{16}$$

在匹配条件下，要超越随机机会，就需要从下一个潜变量中获取除 H 之外的信息。下面的结果明确展示了其与回归的对比。

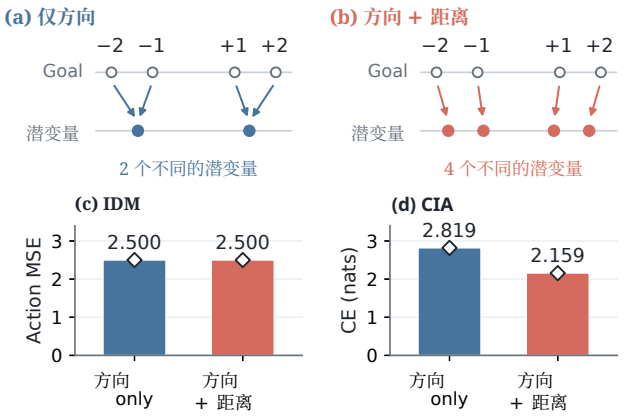


图 2。**超越条件分布动作均值的奖励信息**。上：固定编码仅保留方向，或同时保留方向和距离。下：加入距离不会改变学习值 IDM 均方误差，但会降低 CIA 损失。柱状图表示五次初始化后的平均测试集损失；菱形表示理论最优值。

推论 1（超越条件均值的动作信息）。在命题 *1* 和 *2*的假设下，固定编码器和 H。如果

$$\mathbb{E}[A \mid H, Z'] = \mathbb{E}[A \mid H] \quad \textit{almost surely},$$

但是 $I(A;\ Z' \mid H) > 0$，那么

$$V_H - V_{H,Z'} = 0, \quad \inf_S \mathcal{L}_{\mathrm{CIA}}(S; p_\beta) < \log K, \tag{17}$$

其中 $K \geq 2$且评分取遍不受限制的可测候选函数。

推论1 为 CIA 保留动作信息提供了更宽泛的判据：即使条件分布动作的均值保持不变，其变化仍可能降低最优对比损失。证明见附录 B.3。下面我们通过向一个已经保留方向的表征中添加距离信息，来说明同样的区别。

方向–距离示例。我们用两种固定编码来说明这种对比。从一个固定起始状态出发，动作是 $A = (D,\ B)$。D 在 $\{-1, 1\}$上均匀分布，B 则独立地在 $\{-2, -1, 1, 2\}$上均匀分布。智能体最终到达 $X' = D|B|$，因此 D 决定方向，$|B|$ 决定距离。一种编码仅保留方向，另一种同时保留方向和距离；二者的方差均为 1。给定方向后，距离能揭示 $|B|$，但不能揭示 B的符号，因此它会改变 B 的分布，却不会改变其均值。因此，两种编码给出的条件动作均值相同，均为$(D,\ 0)$，最优均方误差也相同，均为 $5/2$。在平方误差回归中，保留距离毫无益处，即使

表1。**目标条件规划成功率（%）**。我们报告了五个训练随机种子下的均值和标准误，每个随机种子和规划器使用150个测试任务。最后一行报告了每种方法在全部十个环境–规划器组合上的未加权均值。每行中，方法均值最高和第二高的结果分别以粗体和下划线显示。

环境 规划器		LeWM	TS	CGS	SMWM	AC-MTM	CIA	CIA+CGS
PushT	MPPI	65.3 ± 1.5	59.5 ± 2.8	71.3 ± 1.3	76.9 ± 1.2	80.0 ± 0.8	$\underline{81.9} \pm 1.1$	$\mathbf{82.1} \pm 1.0$
	CEM	88.9 ± 0.7	85.7 ± 1.4	89.5 ± 0.6	88.3 ± 1.5	89.2 ± 1.1	$\underline{91.1} \pm 0.5$	$\mathbf{92.9} \pm 0.4$
Reacher	MPPI	59.6 ± 0.7	62.5 ± 1.2	63.5 ± 0.4	65.2 ± 0.6	63.6 ± 1.3	$\underline{67.3} \pm 1.0$	$\mathbf{68.9} \pm 2.5$
	CEM	78.8 ± 0.1	77.2 ± 0.7	78.8 ± 1.3	80.4 ± 1.2	$\underline{80.5} \pm 0.8$	$\mathbf{82.0} \pm 0.8$	76.8 ± 1.9
Cube	MPPI	61.7 ± 0.5	70.3 ± 0.9	$\underline{75.9} \pm 0.6$	68.1 ± 0.7	74.5 ± 0.2	74.7 ± 2.3	$\mathbf{78.9} \pm 0.5$
	CEM	63.6 ± 0.3	67.1 ± 1.1	$\underline{77.2} \pm 0.3$	69.2 ± 0.4	73.2 ± 1.2	72.5 ± 2.7	$\mathbf{78.8} \pm 0.9$
手指	MPPI	14.8 ± 1.0	15.1 ± 0.8	13.2 ± 0.4	$\mathbf{19.1} \pm 0.9$	17.3 ± 1.0	16.7 ± 0.9	$\underline{18.0} \pm 1.0$
	CEM	28.4 ± 0.7	31.3 ± 2.1	30.1 ± 0.1	27.7 ± 2.0	26.9 ± 2.0	$\mathbf{43.2} \pm 1.1$	$\underline{40.0} \pm 1.6$
Ball-in-Cup	MPPI	16.1 ± 0.6	14.5 ± 0.9	17.7 ± 1.3	15.3 ± 0.8	12.7 ± 0.8	$\mathbf{20.7} \pm 1.6$	$\underline{19.6} \pm 1.3$
	CEM	38.1 ± 1.6	38.4 ± 3.4	34.7 ± 2.3	29.2 ± 1.6	35.1 ± 2.0	$\mathbf{58.0} \pm 2.0$	$\underline{54.1} \pm 2.6$
总体		51.5	52.2	55.2	53.9	55.3	$\underline{60.8}$	**61.0**

它能将近处目标与远处目标区分开来。这正是推论1的模式，只不过这里比较的是两种编码，而非比较 H 与 $(H,\ Z')$ (附录 C.1)。

我们在每种编码上训练 IDM 头并计算 CIA 分数，从均匀行为策略中抽取 32 个备选项；在这里，该策略就是匹配提议分布。加入距离后，学习到的 IDM 均方误差保持不变，但 CIA 交叉熵降低，接近约为 2。82 和 2。16 纳特的理论最优值（图 2）。加入距离后，评分可以拒绝方向正确但幅度错误的备选项。因此，CIA 会奖励能够区分远近目标的信息，而回归不会。附录 C给出了推导、训练细节以及单独的联合学习检查。

学习提议分布。 对于提议分布而言，在没有下一个潜变量的情况下可获得的最大识别增益为

$$b(q) := \log K - \inf_{s(H,A)} \mathcal{L}_{\mathrm{CIA}}(s; q). \tag{18}$$

修正界为

$$I(A; Z' \mid H) \geq \log K - \mathcal{L}_{\mathrm{CIA}}(S; q) - b(q). \tag{19}$$

该修正项 $b(q)$ 考虑了由提议分布引起的识别问题，而无需 Z'。为了使历史条件化替代项符合行为，我们最小化

$$\mathcal{L}_q(\eta) := -\mathbb{E}_{H,A} \log q_\eta\big(A \mid \mathrm{sg}(H)\big). \tag{20}$$

对于固定的历史特征，最小化该目标等价于最小化行为策略在提议分布族内的期望前向 KL 散度

。该 KL 散度构成 $b(q)$ (附录 B.4）的上界，从而提供了一种有原则的拟合目标，而无需在训练过程中进行精确匹配。我们使用单独的优化器，并将采样得到的备选项从计算图中分离。

世界模型训练。 我们保留预测损失和潜变量正则化：

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{WM}} + \lambda \mathcal{L}_{\mathrm{CIA}}. \tag{21}$$

CIA 直接更新编码器和评分；预测器通过预测损失进行学习。算法 1 总结了更新过程；实现细节见附录E.3。

5. 实验

我们首先评估目标条件规划性能，以及学习到的提议与评分的质量。随后，我们分析哪些目标组成部分促成了性能提升，并通过受控消融检验替代动作来源的贡献。

5.1. 规划评估设置

我们评估了五种具有连续动作的基于图像的控制环境：PushT [4], OGBench-Cube [16], 以及 Reacher、Finger（简单转动）和来自 DeepMind Control Suite 的 Ball-in-Cup（接球）[22]。任务要求到达视觉目标配置；环境、数据集和成功的定义见附录D。

我们比较了 LeWM [14], 时间拉直（TS） [25], 控制几何拉直（CGS） [6],SMWM 式逆回归 [10], AC-MTM 对比式逆学习 [3], 和 CIA。所有方法均使用

LeWM 主干网络。尽管它们的原始表述不需要 SIGReg，但我们在基于 LeWM 对 AC-MTM 和 SMWM 的改编版本中保留了该方法，以便在相同的防坍缩正则化下比较辅助目标。我们还评估了 CIA+CGS，以考察对比式动作识别与几何正则化如何协同工作。CIA 使用学习得到的历史条件提议分布$q_\eta(a \mid h)$，在 PushT、Cube、Finger 和 Ball-in-Cup 中使用该分布；在 Reacher 中则使用与已知均匀行为分布相匹配的固定均匀提议分布。训练预处理和配置详见附录D和E。

MPPI [26]和 CEM [5] 在每个环境中使用相同的评估协议，对预测的潜在目标代价进行搜索：执行一个包含 25 个基本动作步的规划，然后重新规划一次，最多执行 50 步。每种方法使用五个训练随机种子，并且对于每个随机种子和规划器使用相同的 150 个测试任务。表1报告跨随机种子的平均值和标准误。

5.2. 目标条件规划性能

如表 1所示，CIA 本身就展现出较强的规划性能：在每种环境–规划器组合中，其平均成功率均高于 LeWM；在对环境和规划器进行等权重处理时，汇总增益达到 $9.3\%p$ 。这些结果支持在潜变量预测中加入动作识别监督和防坍缩正则化。无论使用 MPPI 还是 CEM，都能获得这些增益，且无需改变规划算法或潜在目标代价。

与 SMWM 的比较凸显了对比动作识别的实际价值：CIA 获得了更高的总体规划器平均成功率，其中在手指和 Ball-in-Cup 上的增益最大。在精确行为匹配下，该目标可以奖励超越条件动作均值的区分信息（推论 1）。当这些区分信息由共享编码器保留时，它们可以为前向预测和目标比较提供信息，这可能解释了下文诊断所考察的规划增益。为控制逆向输入的影响，我们还使用 CIA 的历史和查询构造评估了 SMWM-H；CIA 的总体平均值仍高于 SMWM-H，而后者与 CIA 使用相同的逆向输入和查询构造。最明显的增益出现在 Ball-in-Cup 和手指上；环境特定的结果及自助法区间见附录 H。

AC-MTM 提供了一个比较，因为它同样通过对比式逆动作识别来训练表征。CIA 在四个环境中取得了更高的规划器平均值，其中在 Finger 和 Ball-in-Cup 上的优势最为明显。AC-MTM 使用批内替代项，而 CIA 则基于相同的历史抽取候选动作。这种构造旨在强调同一行为上下文中的动作—结果兼容性。这些方法在得分参数化和候选数量方面也有所不同；

表 2. **提议与得分质量**。 验证集为三个随机种子的均值。边界率以 q/数据表示；其他 z' 使用来自另一个查询的下一个潜变量。交叉熵的单位为纳特。

	提议 q_η		评分 S_ψ		
环境	样本 var.	边界 (%)	Acc. (%)	CE	其他 z' 准确率 (%)
PushT	0.19	0.03/ 0.02	57.6	1.49	4.3
Cube	0.24	7.2 / 7.4	89.0	0.34	3.9
Ball-in-Cup	0.25	0.8 / 0.6	89.7	0.27	2.9
手指	0.61	11.2/ 10.9	80.1	0.51	2.3

第 5.5 在 CIA 中隔离候选源效应。

在 Cube 上，CIA 取得了具有竞争力的表现，并且增加 MPPI 采样预算后，其与 CGS 的差距有所缩小。对相同冻结模型的诊断表明，将几何正则化与动作识别结合起来，或许能够进一步改善规划（附录I）。CIA+CGS 在 Cube 和 PushT 上取得了最高的规划器平均值，并且总体平均值略高于单独使用 CIA 的结果（61。0%，而 60。8%）。单独使用 CIA 时，在 Reacher、Finger 和 Ball-in-Cup 上仍保持更高的平均值。组合配置使用了分别指定的权重，因此这些结果展示的是组合的潜力，而非孤立的加性效应，也并不意味着在所有环境中都能获得均匀的收益。

5.3. 学习到的提议与评分的质量

提议分布提供替代动作，而评分模型则学习利用历史和下一个潜变量识别已执行动作。训练结束时，我们在四个环境中使用学习到的提议分布，并在三个训练随机种子的留出验证窗口上评估两者；Reacher 使用已知的均匀行为采样器。

学习到的提议分布提供了适应行为数据的备选项。其样本的历史内离散度有限，并且大致复现了整体动作边界频率（表 2）。样本方差是对训练集标准化动作坐标中历史内候选方差的平均值。似然拟合使这些候选项适应条件行为；第 5.5 节和附录 F中的新探针用于衡量缺少下一个潜变量时仍然存在的识别线索。

训练得分利用下一个潜变量识别已执行的动作：在57.6–89.7% 的 $K = 33$候选项中，准确率为此；若替换下一个潜变量，准确率则降至约3.03%的随机机会（表2）。这证实学习得到的评分会按 CIA 的设计，对观测转移作出响应。第5.5 节考察备选项的选择如何影响动作识别及后续规划性能。

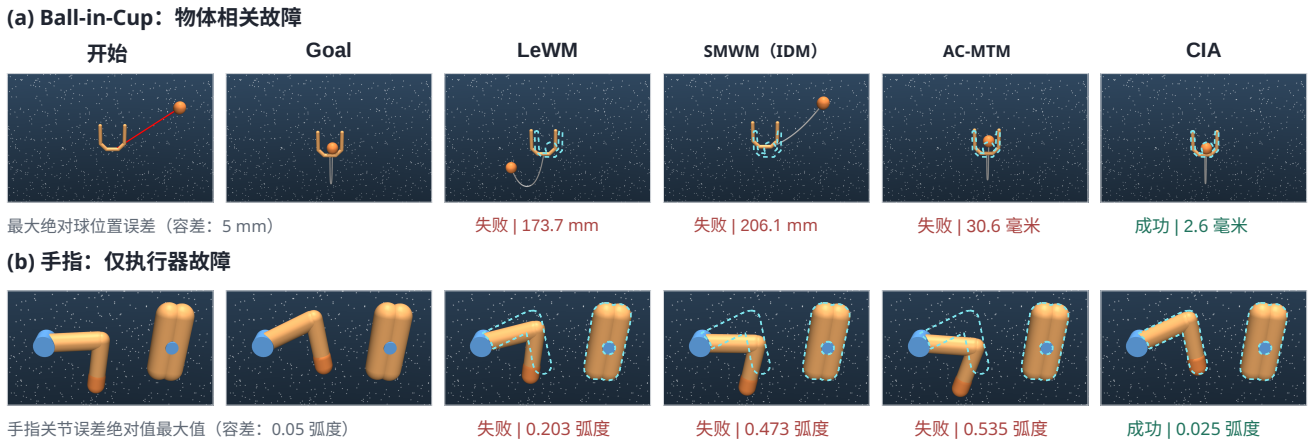


图 3. **规划失败示例**。Ball-in-Cup 展示了 25 个 MPPI 步骤后的状态，而 Finger 展示了 50 个 CEM 步骤后的状态。LeWM、SMWM 和 AC-MTM 未能达到球位置目标（上）或手指关节目标（下），而 CIA 成功完成了任务。青色虚线轮廓表示目标构型。

5.4. 规划失败分析

为了确定哪些目标要求得到改善，我们对共享测试任务中未成功的滚动执行进行分类。在从未完全成功的回合中，若物体满足其目标但受驱动组件未满足，则将最终状态归类为仅执行器故障；若物体未达到目标，则归类为物体相关故障，其中包括关节故障。这些互斥类别以所有回合的百分比表示。

表3显示，在 PushT 上，三个逆向目标都能降低仅推杆故障率，而 CIA 还具有最低的与方块相关的故障率。因此，这些故障类别区分了执行器定位方面的改进与实现物体目标方面的改进。相较于 AC-MTM，CIA 的额外降幅较小，并且在重采样条件下存在不确定性。Ball-in-Cup 提供了更清晰的区分：CIA 将与球相关的故障降低了$15.9\%p$ 相较于 AC-MTM，而仅杯子故障十分少见。大多数评估任务开始时球已在杯子内；该结果主要涉及将球保持在杯中并运送至其目标位置，而非一般意义上的接球能力。

手指任务呈现出类似的模式：CIA 主要减少了旋转器达到目标、但手指关节未达到目标的失败，而与旋转器相关的失败变化较小。在这里，即使物体达到目标构型，目标仍有另一部分未满足：手指也必须达到指定的关节构型。这与 Ball-in-Cup 形成对比，后者减少最多的是与球相关的失败。因此，在不同环境中，性能提升涉及目标中不同的未满足部分，而非始终偏向物体或执行器。这些组成部分并未被设为独立的训练目标；这种分解表明，在不同环境中观察到的规划提升分别对应于不同的目标要求。图 3展示了体现这两种模式的部分匹配推演。

表 3. **失败分解**。所有回合所占的百分比，对 MPPI 和 CEM 以及五个随机种子取平均。类别依据从未成功的回合的最终状态划分。数值越低越好。

环境	故障类型	LeWM	TS	CGS	SMWM	AC-MTM	CIA
PushT	仅推块器	9.0	13.5	6.4	1.8	1.8	2.5
	方块	13.9	13.9	13.2	15.6	13.6	11.0
手指	仅关节	39.1	40.8	41.8	38.7	39.5	33.5
	旋转器	39.3	36.0	36.5	37.9	38.3	36.6
Ball-in-Cup	仅杯子	0.7	0.5	1.0	0.8	0.5	1.0
	Ball	72.1	73.0	72.8	76.9	75.6	59.7

示例。

冻结编码器读出为 PushT 提供了另一项观察结果（图1 及附录 G）。我们对每个冻结编码器采用相同的线性读出协议，以测试无需调整表征即可恢复目标组成部分的难易程度。在三个随机种子下，与 SMWM 和 AC-MTM 相比，CIA 在准确保留推动器位置的同时，能更准确地呈现方块朝向。逆向方法的推动器位置误差中位数为 3–6 像素，而 LeWM 为 20–22 像素；CIA 的朝向误差中位数也达到与 LeWM 相当的 4–5 度。因此，它兼具准确的执行器读出和 LeWM 所保留的物体方向读出。PushT 目标判据要求这两个组成部分缺一不可。这些读出结果表明目标组成部分是可获取的，但并未直接检验条件动作均值之外的动作信息。

5.5. 替代动作源的影响 s

候选项的构造决定了哪些动作差异用于监督编码器。精确行为匹配为历史条件化替代项提供了依据（命题2）；在此，我们使用从离线数据中学习得到的提议分布，测试其实际效果。在 Ball-in-Cup、Cube 和 Finger 上，我们重新训练

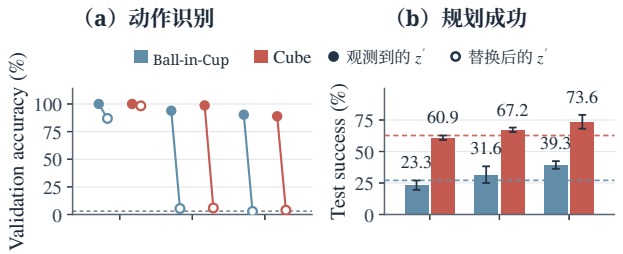


图4。**Ball-in-Cup 和 Cube 上的替代动作来源**。（a）使用观测到的（实心）或替换的（空心）下一个潜变量时的识别准确率。（b）MPPI–CEM 的平均测试成功率。结果展示了五个随机种子的平均值；（b）中的误差线表示跨随机种子的标准误。虚线分别表示（a）中的随机机会和（b）中每个环境的 LeWM 平均值。

CIA，使用从以下三种来源之一抽取的 32 个备选项：动作空间上的均匀分布 $q_{\text{unif}} = \text{Unif}(\mathcal{A})$，它不考虑行为；$M$ 训练动作的经验边际分布 $q_{\text{marg}} = \frac{1}{M}\sum_{i=1}^{M}\delta_{a_i}$，它保留总体动作分布；以及学习得到的提议分布 $q_{\text{cond}} = q_\eta(\cdot \mid h)$，它还会考虑当前历史。

历史条件化替代项在全部三个环境中都取得了最高的规划器平均成功率。在 Ball-in-Cup 和 Cube 中，自助法重采样表明，它们优于边缘替代项（表 4 和附录J）。Finger 的平均值排序也相同，但条件分布与边缘分布差距较小且不确定；均匀替代项在该环境中的表现也优于 LeWM。这些结果支持了拟合提议分布中的历史条件化替代项所具有的实用价值，不过其益处因环境和规划器而异。

图 4 区分了识别难易程度与规划性能。均匀替代项的识别准确率最高，但在三种来源中，平均规划成功率最低。固定历史和候选项、替换下一个潜变量后，其准确率仍然很高；而条件得分则降至接近随机机会的水平，边际得分仍保持略高于随机机会的准确率。尽管识别准确率较低，条件替代项的平均规划成功率仍高于边缘替代项。这些候选项集合定义了不同的分类任务：更容易解决的任务未必能提供更有用的表征监督。

不含下一个潜变量的动作识别。重新放置测试考察各变体的训练得分。在 Ball-in-Cup 和 Cube 上，我们分别重新拟合不使用 z' 的判别器，仅使用候选项（Cand.）、物理状态历史（Phys.）或冻结的 CIA 潜变量历史（潜变量）。这些探针使用相同的条件分布-CIA 模型和数据划分，并对每种来源的三次探针

表4。**替代动作来源**。MPPI–CEM 测试成功率是在五个随机种子上的均值 ± 标准误。$\widehat{\Delta}_X$ 用于衡量不含下一个潜变量时的交叉熵增益。

环境来源		成功率（%）	$\widehat{\Delta}_X$（纳特） 候选物理	潜变量	
Ball-in-Cup	LeWM	27.1 ± 0.8	–	–	–
	均匀	23.3 ± 1.6	3.49	3.49	3.49
	边缘分布	31.6 ± 2.9	0.03	2.27	1.82
	条件分布 q_η	**39.3** ± 1.3	0.90	1.88	0.99
Cube	LeWM	62.7 ± 0.4	–	–	–
	均匀	60.9 ± 0.8	3.50	3.50	3.50
	边缘分布	67.2 ± 0.8	0.03	3.33	3.32
	条件分布 q_η	**73.6** ± 2.5	1.27	1.42	1.22

次初始化，并在取决于模型的情况下，使用三个编码器随机种子。我们报告其相对于随机机会的留出交叉熵增益，

$$\widehat{\Delta}_X = \log 33 - \mathcal{L}_X$$

（表4）。增益越大，表示在不使用下一个潜变量的情况下可用于识别的线索越多；诊断详情见附录F。

与边缘采样相比，条件采样降低了两种环境中测得的潜变量-历史交叉熵增益；在 Cube 中，该数值从 3。32 降至 1。22纳特。仅候选线索仍然存在，并且比使用边缘备选项时更强（附录 F）。因此，拟合提议分布改变了可用的识别线索，而不是将其消除。结合规划消融实验，这些结果支持了学习区分从同一历史中采样得到的动作的实际价值，这一点正如公式（14）中匹配参考情形所启发的那样。

6. 结论与局限性

我们提出了 CIA，这是一种辅助目标，用于从历史条件化提议中采样的备选项中识别已执行的动作。尽管来自下一个潜变量的额外信息只有通过改变条件动作均值，才能改善最优平方误差回归，但即使该均值保持不变，在精确行为匹配下，它仍然可以降低最优对比识别损失。在五个视觉控制基准上，无需修改规划器，CIA 取得了高于 LeWM 和回归 IDM 基线的规划器平均成功率。这些结果表明，采用历史条件化替代项进行对比动作识别，是离线世界模型学习的一种有用目标。

CIA 取决于其条件动作提案的质量。行为匹配不完善可能会留下线索，使模型无需使用下一个潜变量即可识别已执行动作，而似然拟合不一定能消除这些线索。此外，CIA 学习的是一个独立评分，该评分在规划过程中并不使用。其动作识别目标并未直接优化多步预测准确率。

大语言模型使用声明

我们使用 OpenAI ChatGPT、OpenAI Codex 和 Anthropic Claude Code 进行研究构思和理论讨论、代码开发、调试和实验自动化、分析与图表制作，以及语言编辑。图表根据实际实验数据生成，所用绘图代码是在大语言模型的辅助下开发的。作者对科学论断、实验结果、引文及所有提交内容负责。

参考文献

[1] Mahmoud Assran、Adrien Bardes、David Fan、Quentin Gar- rido、Russell Howes、Mojtaba Komeili、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zho- lus、Sergio Arnaud、Abha Gejji、Ada Martin、Francois Robert Hogan、Daniel Dugas、Piotr Bojanowski、Vasil Khalidov、Patrick Labatut、Francisco Massa、Marc Szafraniec、Kapil Krishnakumar、Yong Li、Xiaodong Ma、Sarath Chandar、Franziska Meier、Yann LeCun、Michael Rabbat 和 Nico- las Ballas。V-JEPA 2：自监督视频模型实现理解、预测和规划。arXiv 预印本 arXiv:2506.09985，2025。1，2[2] Randall Balestriero 和 Yann LeCun。LeJEPA：无需启发式方法的可证明且可扩展的自监督学习。arXiv 预印本 arXiv:

2511 08544, 2025 1, 2, 7 。。Jack Boylan 和 [3]
Chris Hokamp。无需高斯分布：JEPA 世界模型的对比逆动力学。arXiv 预印本 arXiv:2608.17542，2026。1，2，5，8，9[4] Cheng Chi、Zhenjia Xu、Siyuan Feng、Eric Cousineau、Yilun Du、Benjamin Burchfiel、Russ Tedrake 和 Shuran Song。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。2，5，7[5] Pieter-Tjerk de Boer、Dirk P. Kroese、Shie Mannor 和 Reuven Y. Rubinstein。交叉熵方法教程。《运筹学年鉴》，134 67，2005 1 19 6 ( ): – 。[6] Ziang Fu 和 Ning Ning。用于基于采样的潜在规划的控制几何拉直。arXiv 预印本 arXiv:2609.35603，2026。5，8，9[7] Ziang Fu 和 Ning Ning。用于基于采样的潜在规划的控制几何拉直，2026 1, 2 。

David Ha 和 Jurgen [8] Schmidhuber。世界模型。arXiv 预印本 arXiv:1803.10122，2018。1[9] Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。软演员-评论家：采用随机演员的离策略最大熵深度强化学习。发表于第 35 届国际机器学习会议论文集，页码 1861 1870 2018 7 –。

PMLR，。Petr Ivashkov、Randall Balestriero 和 Bernhard [10] Sch¨olkopf。感觉运动世界模型：通过逆动力学实现面向行动的感知。arXiv 预印本 arXiv:2606.20104，2026。1，2，3，5，8，9，12[11] Thomas Kipf、Elise van der Pol 和 Max Welling。结构化世界模型的对比学习。发表于国际学习表征会议，2020。2 [12] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于《国际学习表征会议》，2019 年。9[13] Zhuang Ma 和 Michael Collins。条件模型的噪声对比估计与负采样：一致性与统计效率。载于《2018 年自然语言处理经验方法会议论文集》，第 3698–3707 页，比利时布鲁塞尔，2018 年。计算语言学协会。4, 2[14] Lucas Maes、Quentin Le Lidec、Damien Scieur、Yann LeCun 和 Randall Balestriero。LeWorldModel：基于像素的稳定端到端联合嵌入预测架构。arXiv 预印本 arXiv:2603.19312，2026 年。1, 2, 3, 5, 7, 8, 12[15] Tung D Nguyen、Rui Shu、Tuan Pham、Hung Bui 和 Stefano Ermon。用于潜在空间中基于模型的规划的时间预测编码。载于《第38届国际机器学习会议论文集》，第 8130–8139 页。PMLR，2021 年。2[16] Seohong Park、Kevin Frans、Benjamin Eysenbach 和 Sergey Levine。OGBench：离线目标条件强化学习基准。载于《国际学习表征会议》，2025 年。2, 5, 7[17] Keiran Paster、Sheila A. McIlraith 和 Jimmy Ba。使用逆动力学模型进行基于像素的规划。载于《国际学习表征会议》，2021 年。2[18] Ben Poole、Sherjil Ozair、Aaron van den Oord、Alex Alemi 和 George Tucker。关于互信息的变分界。载于《第36届国际机器学习会议论文集》，第 5171–5180 页。PMLR，2019 年。4[19] Kate Rakelly、Abhishek Gupta、Carlos Florensa 和 Sergey Levine。哪些互信息表征学习目标足以实现控制？载于《神经信息处理系统进展》，2021 年。2[20] Vlad Sobal、Wancong Zhang、Kyunghyun Cho、Randall Balestriero、Tim G. J. Rudner 和 Yann LeCun。从无奖励离线数据中学习：基于潜在动力学模型进行规划的案例。arXiv 预印本 arXiv:2502.14819，2025 年。1[21] Alessandro Sordoni、Nouha Dziri、Hannes Schulz、Geoff Gordon、Philip Bachman 和 Remi Tachet Des Combes。用于对比表征学习的分解式互信息估计。载于《第38届国际机器学习会议论文集》，第 9859–9869 页。PMLR，2021 年。1, 4[22] Yuval Tassa、Yotam Doron、Alistair Muldal、Tom Erez、Yazhe Li、Diego de Las Casas、David Budden、Abbas Abdolmaleki、Josh Merel、Andrew Lefrancq 等。DeepMind Control Suite。arXiv 预印本 arXiv:1801.00690，2018 年。5, 7[23] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。使用对比预测编码进行表征学习。arXiv 预印本 arXiv:1807.03748，2018 年。4 [24] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。使用对比预测编码进行表征学习。arXiv 预印本 arXiv:1807.03748，2018 年。2, 4[25] Ying Wang、Oumayma Bounou、Gaoyue Zhou、Randall Balestriero、Tim G. J. Rudner、Yann LeCun 和 Mengye Ren。用于潜在规划的时间拉直。载于《

第*43*届国际机器学习会议论文集，第 125168–125196 页。PMLR，2026 年。 1， 2， 5[26] Grady Williams、Andrew Aldrich 和 Evangelos A. Theodorou。模型预测路径积分控制：从理论值到并行计算。《制导、控制与动力学杂志》，40(2):344–357，2017 年。 6[27] 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实现零样本规划。载于《第*42*届国际机器学习会议论文集》，第 79115–79135 页。PMLR，2025 年。 1， 2

通过动作的对比识别学习世界模型补充材料

A. 回归逆动力学分析

以下结果涉及一个固定编码器和指定的历史变量 H，假设 $\mathbb{E}\|A\|_2^2 < \infty$。对预测器取的所有下确界，均为在具有有限风险的全体可测函数中取得的总体量。

A.1. 命题 1的证明

证明。 记 $m_H = \mathbb{E}[A \mid H]$ 和 $m_{HZ} = \mathbb{E}[A \mid H, Z']$。条件期望给出正交分解

$$\begin{aligned}\mathbb{E}\|A - r(H, Z')\|_2^2 = \mathbb{E}\|A - m_{HZ}\|_2^2 \\ + \mathbb{E}\|m_{HZ} - r(H, Z')\|_2^2,\end{aligned} \tag{22}$$

因为给定 $(H,\ Z')$ 时，交叉项的条件期望为零。因此，最小化预测器为 m_{HZ}，从而得到$V_{H,\ Z'} = \mathbb{E}\|A - m_{HZ}\|_2^2$。同理， $V_H = \mathbb{E}\|A - m_H\|_2^2$。令 $r(H,\ Z') = m_H$，代入上述分解可得

$$V_H = V_{H,Z'} + \mathbb{E}\|m_{HZ} - m_H\|_2^2, \tag{23}$$

这证明了（11）。由于右侧的两项均为非负数， $0 \le \Delta \le V_H$。对于 $V > 0$，定义 R_H^2 =1 $- V_H/V$得到 $V_H =$（$1 - R_H^2$）V。等价地，动作总方差可分解为

$$V = R_H^2 V + \Delta + V_{H,Z'}, \tag{24}$$

由历史　　解释的部分、 Z'带来的额外贡献，以及剩余的预测误差。 □

A.2. 解释与限定条件

匹配条件变量。命题 1 使用相同的历史 H比较两个回归问题。如果逆向头仅接收受限的历史摘要 $\chi(H)$以及 Z'，那么其仅基于历史的基线由 $\chi(H)$ 而非 H 决定，并且其从观测 Z' 中获得的增益可能超过 Δ。例如，如果 A 由 H决定，χ为常数，并且 Z'揭示了 A，那么即使将 Z' 加入 $\chi(H)$ 会使最小误差从 V降至零，对于完整历史而言， $\Delta = 0$ 。

均值保持的动作信息。 根据（11）， $\Delta = 0$ 当且仅当

$$\mathbb{E}[A \mid H, Z'] = \mathbb{E}[A \mid H] \quad \text{almost surely.} \tag{25}$$

然而，条件均值相等并不意味着条件分布相等。令 H 为常数， A 在$\{-2, -1, 1, 2\}$上均匀分布，并令 $Z' = |A|$。给定 Z'的任一取值，两个可能的动作符号相反且概率相等，因此

$$\mathbb{E}[A \mid H] = \mathbb{E}[A \mid H, Z'] = 0, \qquad V_H = V_{H,Z'} = \frac{5}{2}. \tag{26}$$

然而， Z' 能够准确揭示动作幅度：记香农熵为 H,

$$I(A; Z' \mid H) = \mathsf{H}(Z' \mid H) - \mathsf{H}(Z' \mid H, A) = \log 2. \tag{27}$$

平方误差回归不会因这种信息而受益，而推论1以及第B.3 节中的计算表明，行为匹配对比学习则会。

范围。 本分析专门针对在固定编码器和历史 H下的平方误差逆回归；概率逆模型还可以对超出条件均值的分布变化作出响应。CIA 中行为匹配的独特作用，是在理想提议分布下消除仅基于历史的判别。由于这些结果处于总体层面且表征固定不变，因此它们无法直接刻画有限样本优化、编码器梯度或下游规划。第 K 节通过在冻结的基准编码器上测量仅基于历史的可预测性和下一潜变量回归增益，为本分析提供了实证补充。

B. 条件逆对比学习分析

固定编码器和条件变量 H。我们假设存在正则条件分布和共同的动作空间参考测度（其中可能包括删失动作的边界原子），在该测度下，$p_\beta(\cdot \mid h,\ z')$、$p_\beta(\cdot \mid h)$ 和 $q(\cdot \mid h)$ 均具有密度，并且 $p_\beta(\cdot \mid h) \ll q(\cdot \mid h)$ 几乎处处成立。备选项从 $q(\cdot \mid H)$ 中条件独立地抽取，并且在给定 H 的条件下，与 $(A_0,\ Z')$ 独立。所有对数均采用自然对数，且 $K \geq 2$ 统计所有候选项。对评分取下确界时，范围涵盖不受限可测逐候选函数；当后验概率为零而需要发散的逻辑值时，也允许采用极限序列。

B.1. 随机索引表示与最优评分

设 J 在 $\{0, \ldots, K-1\}$ 上均匀分布，且与 $(H、\ A_0、\ Z'、\ A_1、\ \ldots\ldots、\ A_{K-1})$ 相互独立。通过将已执行动作置于 $C_J = A_0$，并将备选项置于其余位置，构造候选元组 C $=(C_0、\ \ldots\ldots、\ C_{K-1})$。对于共享的逐候选项评分 S，定义

$$\pi_S(j \mid h, z', c) = \frac{\exp S(h, z', c_j)}{\sum_i \exp S(h, z', c_i)}. \tag{28}$$

由于同时置换候选项及其标签不会改变损失，

$$\mathcal{L}_{\mathrm{CIA}}(S; q) = \mathbb{E}[-\log \pi_S(J \mid H, Z', C)]. \tag{29}$$

Let

$$\rho(h, z', a) = \frac{p_\beta(a \mid h, z')}{q(a \mid h)}. \tag{30}$$

在给定 $(H、\ Z'、\ J)$ $=(h、\ z'、\ j)$ 的条件下，候选项密度为 $p_\beta(c_j \mid h、\ z') \prod_{i \neq j} q(c_i \mid h)$。根据贝叶斯定理，

$$\pi^*(j \mid h, z', c) := \mathbb{P}(J = j \mid h, z', c) = \frac{\rho(h, z', c_j)}{\sum_i \rho(h, z', c_i)}. \tag{31}$$

记 H 为香农熵，交叉熵可分解为

$$\begin{aligned} \mathcal{L}_{\mathrm{CIA}}(S; q) = {} & \mathsf{H}(J \mid H, Z', C) \\ & + \mathbb{E}\,\mathrm{KL}\big(\pi^*(\cdot \mid H, Z', C) \,\|\, \pi_S(\cdot \mid H, Z', C)\big). \end{aligned} \tag{32}$$

因此，$S^* = \log \rho + c(h, z')$ 在有限处均为贝叶斯最优，并且可确定到一个与动作无关的加性项 $c(h, z')$。将 $p_\beta(a \mid h, z') = p(z' \mid h, a) p_\beta(a \mid h) / p(z' \mid h)$ 代入即可证明（14）。其中，$p(z' \mid h, a)$ 是数据分布下的条件密度；如果 h 无法确定完整状态，它仍可能通过未观测状态依赖于行为策略。当后验为某些候选项分配零概率时，最优解需要 $-\infty$ logits，而有限的神经评分可以在极限情况下逼近这一结果。这恢复了对比分类的标准密度比解释 [13, 24]。

B.2. 命题 2 的证明

证明。 利用第 B.1 节中具有正位置的随机索引构造 J 和候选元组 C，精确匹配（$q = p_\beta$）意味着，对 Z' 进行边缘化后，给定 K，所有 H候选项均独立同分布。因此，J 在给定 $(H,\ C)$时服从均匀分布，并且

$$\mathsf{H}(J \mid H, C) = \log K. \tag{33}$$

任何历史—只有评分的交叉熵至少达到该熵，从而证明（15）。根据（32），每个评分 S 都满足

$$\begin{aligned} \log K - \mathcal{L}_{\mathrm{CIA}}(S; p_\beta) &\leq \mathsf{H}(J \mid H, C) - \mathsf{H}(J \mid H, Z', C) \\ &= I(J; Z' \mid H, C). \end{aligned} \tag{34}$$

由于 $A_0 = C_J$由（$J,\ C$）决定，并且在给定（$H,\ A_0$）的条件下，（$J,\ C$）与 Z' 条件独立，因此我们有

$$I(J, C; Z' \mid H) = I(A_0; Z' \mid H). \tag{35}$$

应用链式法则可得 $I(J;\ Z' \mid H,\ C) \leq I(J,\ C;\ Z' \mid H) = I(A_0;\ Z' \mid H) = I(A;\ Z' \mid H)$，从而证明了（16）。注意，（35）对于任何满足相同条件独立采样假设的一般 q 均成立。 □

B.3. 推论1的证明

我们首先证明，条件依赖等价于严格为正的最优对比式增益，并采用最优得分刻画，而非仅依赖互信息界。

证明。 在精确行为匹配下，定义最优对比增益

$$G_K := \log K - \inf_S \mathcal{L}_{\mathrm{CIA}}(S; p_\beta). \tag{36}$$

由于密度比得分在（31）中实现了贝叶斯后验（以上述极限意义而言）， 下确界$_S\,\mathcal{L}_{\mathrm{CIA}}$（$S$； p_β） $=$H（J |H， Z', C）。结合 H（$J \mid H$， C） $=$对数 K，可得

$$G_K = I(J; Z' \mid H, C). \tag{37}$$

我们声称

$$G_K = 0 \quad \Longleftrightarrow \quad A \perp\!\!\!\perp Z' \mid H. \tag{38}$$

如果 $A \perp\!\!\!\perp Z' \mid H$，则密度比恒等于一，关于 J 的后验是均匀的，并且 $G_K =$ 0.

反之，假设 $G_K = 0$，因此 J 给定 (H、 Z'、 C) 的后验几乎处处均匀。对于几乎所有的 (h、 z')，记 $P_h = p_\beta(\cdot \mid h)$，并且

$$\rho(a) = \frac{p_\beta(a \mid h, z')}{p_\beta(a \mid h)}, \qquad \int \rho(a)\, P_h(\mathrm{d}a) = 1. \tag{39}$$

相对于 $Q = P_h^{\otimes K}$，给定 (h, z') 的候选分布具有密度 $\rho(c) = K^{-1}\sum_i \rho(c_i)$。当 $\rho(c) > 0$ 时，均匀后验要求所有 $\rho(c_i)$ 都相等；当 $\rho(c) = 0$ 时，非负性意味着它们全都为零。因此， $\rho(C_0) = \rho(C_1)$ Q- 几乎处处成立。由于在 Q 下， C_0 和 C_1 独立同分布， ρ 必须是 P_h-几乎处处的常数，而单位期望迫使 $\rho = 1$。因此， $p_\beta(\cdot \mid h, z') = p_\beta(\cdot \mid h)$，从而证明了 (38)。

在推论 1 的假设下，条件均值相等意味着 $\Delta = 0$ 根据命题 1，而 $I(A; Z' \mid H) > 0$排除了条件独立性，因此根据（ $G_K > 0$ （38）。等价地，下确界$_S\,\mathcal{L}_{\mathrm{CIA}}(S;\ p_\beta) <$对数 K。 □

幅值示例中的对比损失。 回到 A.2 节中的示例，A.2，对于 $q = p_\beta$，候选项具有非零后验概率，当且仅当其幅度等于 Z'。所有此类候选项的密度比均为2，其总数（包括已执行的动作）为

$$N = 1 + B, \qquad B \sim \mathrm{Binomial}(K - 1, 1/2). \tag{40}$$

贝叶斯后验在这些 N 候选项上均匀分布，因此得到

$$\begin{aligned} \inf_S \mathcal{L}_{\mathrm{CIA}}(S; p_\beta) &= \mathbb{E}[\log N], \\ G_K &= \log K - \mathbb{E}[\log N] > 0. \end{aligned} \tag{41}$$

对于所有 $K \geq 2$ ，该不等式都是严格的，因为 $N \leq K$ 且 $\mathbb{P}(N < K) > 0$。对于 $K = 2$, - $G_2 = \frac{1}{2}\log 2$, 而回归增益仍为 $\Delta = 0$。从表征中移除幅度后，最优回归误差保持不变，同时对比增益被消除。更丰富的方向–距离示例比较了两种信息性编码，而不是恒定历史，详见 C 节C。

比较的局限性。 推论 1 刻画了固定表征下匹配目标的总体容量。在实践中，有限的神经评分可能无法捕捉所有依赖关系，有限的K 增益上限为 $\log K$ （且不一定能根据互信息对表征进行排序），而不完善的提议分布会引入仅历史偏差 b（q），下一节将对此进行分析。目标相关性和规划效用将在我们的实验中进行实证评估。

B.4. 精确提案失配校正

定义行为先验偏差

$$b(q) := \log K - \inf_s \mathcal{L}_{\mathrm{CIA}}(s; q), \tag{42}$$

其中， $s(h,\ a)$ 遍历忽略 Z' 的不受限评分。这衡量了仅根据历史和候选项所能实现的、低于 $\log\ K$ 的最大降幅。

命题 3（提案失配校正）。 对于每个评分 S,

$$I(A; Z' \mid H) \geq \log K - \mathcal{L}_{\mathrm{CIA}}(S; q) - b(q). \tag{43}$$

此外， $0 \leq b$（q） $\leq \log K$，其中 b（q） $= 0$ 当且仅当 q（$\cdot \mid H$） $= p_\beta$（$\cdot \mid H$） 几乎处处成立 .

证明。 令 $\rho_0(h,\ a) = p_\beta(a \mid h)/q(a \mid h)$。不使用 Z' 重复后验计算可得 s

$$\mathbb{P}(J = j \mid H = h, C = c) = \frac{\rho_0(h, c_j)}{\sum_i \rho_0(h, c_i)}. \tag{44}$$

由于 $\log \rho_0$ 实现了该后验,

$$\begin{aligned} \inf_s \mathcal{L}_{\mathrm{CIA}}(s; q) &= \mathsf{H}(J \mid H, C), \\ b(q) &= I(J; C \mid H). \end{aligned} \tag{45}$$

界 $0 \leq b(q) \leq \log K$ 可直接由 J 的熵推出。在精确匹配下，后验分布是均匀的，且 $b(q) = 0$。反过来，若 $b(q) = 0$，则后验分布是均匀的，并且所有候选项几乎处处具有相同的 ρ_0。在给定 $H = h, J = 0$ 的条件下，独立变量 $C_0 \sim p_\beta(\cdot \mid h)$ 和 $C_1 \sim q(\cdot \mid h)$ 几乎处处具有相同的密度比，因此 $\rho_0(h, \cdot)$ 必为常数；归一化条件 $\int \rho_0(h, a)\, q(\mathrm{d}a \mid h) = 1$ 将该常数确定为 1，因此 $q = p_\beta$ 几乎处处成立。

最后，应用 (35)（它对一般的 q 都成立）可得

$$\begin{aligned} &\log K - b(q) - \mathcal{L}_{\mathrm{CIA}}(S; q) \\ &\quad \leq \mathsf{H}(J \mid H, C) - \mathsf{H}(J \mid H, Z', C) \\ &\quad = I(J; Z' \mid H, C) \leq I(A; Z' \mid H). \end{aligned} \tag{46}$$

□

偏差的 KL 上界。定义

$$D_\beta(q) := \mathbb{E}_H \, \mathrm{KL}\big(p_\beta(\cdot \mid H) \,\|\, q(\cdot \mid H)\big). \tag{47}$$

我们证明 $b(q) \leq D_\beta(q)$。固定 h，令 $P_{j,h}$ 为给定 $H = h, J = j$ 时的候选分布，令 $M_h = K^{-1} \sum_j P_{j,h}$，并令 $Q_h = q(\cdot \mid h)^{\otimes K}$。当散度有限时,

$$\begin{aligned} \mathrm{KL}(p_\beta(\cdot \mid h) \| q(\cdot \mid h)) &= \frac{1}{K} \sum_j \mathrm{KL}(P_{j,h} \| Q_h) \\ &= I(J; C \mid H = h) + \mathrm{KL}(M_h \| Q_h). \end{aligned} \tag{48}$$

Taki对 H 的期望进行..... b建立了 $q(\leq D_\beta)\ q(\ D_\beta)$（当 $q(= \infty)$ 时显然如此），从而得到 s

$$I(A; Z' \mid H) \geq \log K - \mathcal{L}_{\mathrm{CIA}}(S; q) - D_\beta(q). \tag{49}$$

对于固定历史特征和有限期望， $\mathcal{L}_q(\eta) = D_\beta(q_\eta)$ +为常数。因此，对 q_η 进行最大似然拟合会直接最小化提议分布族内仅依赖历史的判别偏差上界。

与现有对比界的关系。 密度比视角和无条件 InfoNCE 界已广为人知 [13,18, 24]。式（49）中的条件化构造和 KL 修正遵循 Sordoni 等人的方法 [21],，将其 (x, y, x') 映射到(Z', A, H)。我们并不主张将这一条件 InfoNCE 界作为新的估计量；相反，我们将其用于比较离线世界模型学习中的逆向目标，在该场景下， $b(q)$ 提供了比 $D_\beta(q)$更紧的操作性修正。

B.5. 解读及与实验的联系

条件化与范围。 上述保证适用于指定的历史 H（在固定编码器下），而不是完整的原始观测历史。提议和评分都使用相同的条件化，并且在给定 H 的情况下，备选项与已执行动作及下一观测彼此独立采样（见第 E.3节）。

解读拟合探针。 提议分布的偏差取决于它与数据收集行为的匹配程度：均匀采样匹配均匀随机行为，而边缘采样匹配与历史无关的行为。对于任意拟合的仅依赖历史的评分 $\widehat{s}$，

$$\log K - \mathcal{L}_{\mathrm{CIA}}(\widehat{s}; q) \le b(q). \tag{50}$$

因此，正的探针增益可以确认存在结果盲判别，而增益较小并不能证明其不存在。由于拟合探针仅能给 $b(q)$ 提供下界（受有限样本方差影响），从训练得分的增益中减去该值，并不能保证得到 $I(A; Z' \mid H)$ 的有效下界。

匹配与拟合的提议分布。 在精确行为匹配下，任何仅依赖历史的评分都无法优于随机机会，因此逆向任务必须依赖 Z'。对于一般提议分布，公式（14）将最优评分分解为转移兼容性和行为先验比，而 $b(q)$ 则量化了仅依赖历史的最大增益。通过最大似然拟合q_η，可以在提议分布族内最小化前向 KL 上界 $D_\beta(q_\eta)$，但这并不保证 $b(q_\eta) = 0$。我们的学习型提议分布实验将这一实用构造与精确匹配的理论参照进行比较（Reacher 通过其已知的均匀采样器直接实现了精确匹配）。

与概率逆动力学的关系。 条件均值局限适用于平方误差回归（如 SMWM），而不适用于概率逆模型；概率逆模型还可以利用均值之外的分布变化（表 5）。然而，CIA 在如何监督表征方面不同于基于似然的逆模型。基于似然的逆模型会奖励所有动作可预测性，包括仅凭历史即可实现的可预测性；在我们的五个环境中的三个环境里，这解释了大部分动作方差（第 K 节）。相比之下，CIA 将基于历史的行为建模交由一个在分离特征上拟合的独立提议，并通过相对识别训练共享编码器；在精确匹配下，要超过随机机会，需要来自 Z' 的信息。此外，CIA 的交叉熵有界，而连续动作的对数似然可能会随着分布集中而发散。针对完整世界模型训练的基于似然的逆向目标进行基准测试，将留待未来工作。

与规划的联系。 我们的分析和实验分别探讨了整体图景中相互补充的部分：推论1 和固定编码器检查证实，动作识别会奖励超越条件均值的信息；候选来源消融检验了历史条件化替代项的实际益处；而状态读出和失败分解则考察了目标分量可及性。总体而言，这些发现表明 CIA 是一种有效的规划辅助监督，尽管基准测试的收益可能同时反映了表征线索、候选覆盖变化以及优化动态的影响。

C. 方向–距离示例

本节详细介绍图2的设置、解析最优解以及固定编码器学习检验，用于测试训练后的神经网络头是否复现 CIA 赋予距离信息的额外奖励。

C.1. 构造与解析比较

固定初始状态和历史。动作是 $A = (D, B)$，其独立分量为 $D \sim \mathrm{Unif}\{-1, 1\}$ 和 $B \sim \mathrm{Unif}\{-2, -1, 1, 2\}$，生成端点 $X' = D|B|$。我们比较两种固定的零均值、单位方差标量编码：

$$Z'_{\mathrm{dir}} = \mathrm{sgn}(X') = D, \qquad Z'_{\mathrm{full}} = X'/\sqrt{5/2}. \tag{51}$$

Z'_{dir} 在每个方向上合并近端点和远端点，而 Z'_{full} 则区分全部四个端点。在任一种编码下，B 的符号仍然同样可能为正或负，因此

$$\mathbb{E}[A \mid Z'_{\mathrm{dir}}] = \mathbb{E}[A \mid Z'_{\mathrm{full}}] = (D, 0), \qquad V_{H, Z'_{\mathrm{dir}}} = V_{H, Z'_{\mathrm{full}}} = 5/2. \tag{52}$$

表 5。**固定编码器学习检验**。学习值为五次初始化的均值；均方误差使用原始坐标，交叉熵/负对数似然的单位为纳特。

编码	指标	学习值	理论值
仅方向	IDM 均方误差	2.500000	2.500000
	CIA 交叉熵	2.819141	2.818752
	概率IDM 负对数似然	1.386317	1.386294
方向 + 距离	IDM 均方误差	2.500000	2.500000
	CIA 交叉熵	2.158911	2.156962
	概率 IDM 负对数似然	0.693300	0.693147

此处，均方误差对两个原始动作坐标误差求和。两种编码都能揭示方向，并将回归误差从 $7/2$降至 $5/2$，从而在方向已知的情况下，单独考察距离带来的增量价值： $\mathbb{E}[A \mid D, |B|] = \mathbb{E}[A \mid D] = (D,\ 0)$，但$I(A;\ |B| \mid D) = \log 2$。不同于推论 1中的零增益标量示例，相对于固定历史的总回归增益并非零。备选项从涵盖全部八种动作的精确均匀行为分布中采样。

对于 CIA，候选集由已执行动作和从均匀行为分布中有放回独立抽取的 $K-1$ 个样本组成（保留重复项），并与正标签一起进行联合置换。兼容的备选项在 Z'_{dir} $(p = 1/2)$ 下方向相同，在 Z'_{full} $(p = 1/4)$ 下方向和幅度均相同。贝叶斯后验在$1 +$ 个兼容候选项上服从均匀分布，其中数量服从二项分布$(K-1,\ p)$，因此最优损失为

$$\ell_K(p) = \sum_{j=0}^{K-1} \binom{K-1}{j} p^j (1-p)^{K-1-j} \log(1+j). \tag{53}$$

通过共享均匀随机变量耦合伯努利抽样可得 $\ell_K\,(1/4) \ < \ell_K\,(1/2)$，对每个 $K \geq 2$均成立。在 $K = 33$时，最优损失分别为 2.818752 和 2.156962 纳特（零后验值按 B.1 节中的极限得分约定处理）。分类概率逆模型同样受益，其最优负对数似然从 $\log 4$ 降至 $\log 2$。

为了将这些编码与规划联系起来，假设从 $\{-2, -1, 1, 2\}$ 中均匀采样一个目标端点，并且只通过其编码提供该端点。在 Z'_{dir} 下，两个距离相同的概率映射到同一表征，因此任何控制器的平均成功率最高只能达到 $1/2$。在 Z'_{full} 下，精确的潜在动力学和穷举动作搜索可实现 100% 的成功率。

C.2. 固定编码器神经学习检验

我们仅在这些固定编码上训练逆向头。独立的回归头、对比头和八分类概率头均使用宽度为 64 的双隐藏层 GELU 多层感知机。回归头和分类头以潜变量 Z' 为输入，而对比评分头以 $(Z',\ a)$ 为输入。我们使用 AdamW 训练（学习率 10^{-3}，权重衰减 10^{-4}，余弦衰减），在五组配对初始化上进行 2,500 次更新，并评估最终检查点，不进行检查点选择。每次更新中，回归均方误差和分类负对数似然都在全部八个状态转移上精确求平均；对比头则每次更新使用 256 个均衡候选集，并采用 $K = 33$。

在评估时，均方误差和分类负对数似然在全部八个状态转移上精确计算，而对比交叉熵则通过 262,144 个独立采样、按状态转移均衡的候选集进行估计；这些候选集在不同编码和初始化之间共享。表 5比较学习到的损失与总体最优值，图2 绘制回归和 CIA 结果。

学习值 CIA 交叉熵与理论值高度吻合，样本标准差为 0。000005 和 0。000276 （小于图 2 中的标记大小）。在同一评估样本库上，预言机交叉熵值为 2。819075 和 2。158121，因此训练得分仅比有限样本库的最优值高出 0。000066 和 0。000791 纳特。总的来说，神经网络头忠实地再现了理论目标差距：回归对距离不敏感，而 CIA 和概率 IDM 都会强烈奖励距离。

C.3. 单独的联合学习检查

固定编码器检查将表征固定不变。这项检查也会学习编码器，因此预测和逆向目标会共同塑造表征。

设置。 观测为 224×224 RGB图像，显示物体和一个独立的视觉干扰项。我们使用 512 个训练集、128 个验证集和 512 个测试视觉上下文。每个上下文都与全部八种动作配对，且干扰项在每次状态转移中保持不变。编码器使用两个步幅为 2、通道数分别为 8 和 16 的 3×3卷积层，后接全



图5. **评估环境**。 从左至右：PushT、Reacher、OGBench-Cube、Finger（转向简单），以及 Ball-in-Cup（接球）。

连接层，其宽度分别为 64 和 8，层间使用 GELU 激活函数。前向预测器和逆向头均为多层感知器，包含两个隐藏层，每层有 64 个 GELU 单元。

目标与训练。 我们比较四种变体。它们都使用预测均方误差和 SIGReg [2]，权重为 0.09，使用 128 个随机单位投影，并在 $[0, 3]$ 上设置 17 个节点。三种变体额外加入一个系数为 0.1 的逆向损失：回归式 IDM、CIA 或概率 IDM（分类负对数似然）。编码器接收所有启用损失的梯度。预测器仅接收预测梯度，且预测目标不停止梯度。我们使用批次大小 128 和 AdamW（学习率 3×10^{-4}，权重衰减 10^{-4}，余弦衰减）训练 3,000 次更新。在五次初始化中，各变体的初始化、数据顺序和正则化随机性均保持配对。

评估与结果。 我们在冻结编码器上使用岭回归探针读取距离，并根据验证集上下文选择正则化参数。一步规划使用学习值预测器枚举全部八种动作，并选择潜在目标距离最小的动作；只有到达指定的精确端点才算成功。在测试集上下文中，所有四种变体在五次初始化中均达到 100% 的规划成功率。距离读取的平均值 R^2 为：使用 SIGReg 的预测模型为 0.9878，回归式 IDM 为 0.9985，CIA 为 0.9882，概率 IDM 为 0.9912。因此，这项检查并未显示 CIA 有优势：在这个简单设置中，共享的预测目标已经能够保留距离信息。目标层面的区别取决于上文的固定编码器检查；这两项检查均未表明，超出条件均值的信息会带来基准任务中的规划收益。

D. 环境和数据集

所有五个环境均使用 224×224 RGB观测和连续动作（图5）。世界模型仅使用不含奖励或任务标签的离线图像—动作轨迹进行训练。环境接口遵循 PushT、 [4], OGBench [16], 和 DeepMind Control Suite [22]。我们使用 LeWM 发布的 PushT 专家、OGBench-Cube 单立方体专家以及 Reacher 随机动作数据集 [14], , 并为 Finger（带噪启发式策略）和 Ball-in-Cup（经过训练的软演员-评论家（SAC）策略 [9]）收集了自己的离线数据集。物理状态仅用于初始化和评估回合评分，以及运行状态诊断。

数据划分和预处理。我们对所有方法使用固定且回合互不重叠的训练集/验证集/测试集划分：PushT 有 16，817/934/934 个回合，Reacher 和 Cube 有 9，000/500/500 个回合，Finger 和 Ball-in-Cup 有 7，200/400/400 个回合。在训练集上计算动作归一化统计量，并在评估时重复使用。记录的 PushT 动作在标准化之前按坐标裁剪至 $[-1, 1]$；所有其他数据集中的动作均已处于 $[-1, 1]$范围内。在规划期间，MPPI 和 CEM 在标准化坐标中优化无约束序列，并将其反标准化后执行。Reacher、Finger 和 Ball-in-Cup 会将已执行动作裁剪至 $[-1, 1]$，而 PushT 和 Cube 则直接将未经标准化的原始动作传递给其环境接口。

对收集到的数据集进行分析。Finger 和 Ball-in-Cup 各自的数据集均包含 8,000 个回合，每个回合包含 200 个原始步（共 1.6M 次状态转移和 1.608M 帧 RGB 帧，分辨率为 224×224），汇总见表 6。正奖励回合指至少有一个步骤获得正环境奖励的回合，这是一项收集数据时的诊断指标，与我们的视觉目标到达评估不同。

表 6。**我们所收集数据集的统计量**。动作标准差是在原始坐标下，针对有限原子动作行计算的，不包括终止填充。

统计量	手指	Ball-in-Cup
回合	8,000	8,000
原子转移	1,600,000	1,600,000
记录帧数	1,608,000	1,608,000
动作标准差（两个坐标）	(0.359, 0.683)	(0。288，0。421)
正奖励回合（%）	53.3	100.0
正奖励最终帧（%）	49.5	100.0

在 Finger 中，我们对启发式策略的每个坐标分别添加独立的高斯噪声（$\sigma = 0.2$），并将已执行动作裁剪至 $[-1, 1]$。累积绝对旋转器转动量（对展开后的角度增量求和）的中位数为每回合 $148.1°$，第 10 和第 90 百分位数分别为 $27.3°$ 和 $245.5°$。

在 Ball-in-Cup 中，我们从基于环境稀疏捕获奖励训练的 SAC策略中随机采样动作（$+1$ 当球位于杯子目标区域内时，奖励为捕获奖励，否则为 0）。每个回合至少获得一次捕获奖励，首次捕获的中位时间为 15 个原始步；95.1% 的非初始步获得捕获奖励，且仅有 5.5% 的回合出现从正奖励到非正奖励的状态转移。因此，捕获后的稳定阶段占据了数据集的大部分。

视觉目标与成功标准。 每个视觉目标到达任务都提供一个初始帧和一个目标帧，这两个帧取自同一条留出轨迹，间隔 24 个原始步（其中一个原始步包含环境内部的任意动作重复）。我们在 200 个验证集任务（四组各 50 个任务）和 150 个测试任务（三组各 50 个任务）上进行评估，所有方法和随机种子均使用相同的任务。在 Finger 和 Ball-in-Cup 上，步骤 0 时已满足成功标准的开始 – 目标对会在采样前被过滤掉。

如果在 50 个基本动作步预算内的任意一步满足环境特定状态标准，则该回合成功：在 PushT 上，拼接后的推块器与方块欧氏位置误差低于 20 px，且环绕方块方向误差低于 $20°$；在 Cube 上，立方体位置误差低于 40 mm；在 Reacher 上，每个关节坐标误差均低于 0。05 弧度；在 Finger 上，每个手指关节和旋转器铰链误差均低于 0。05 弧度；在 Ball-in-Cup 上，杯子与球位置坐标误差均低于 5 mm。DMC 比较使用记录的广义坐标，不进行角度环绕。请注意，失败分解仅使用失败回合的最终状态误差对失败回合进行分类，而不是采用任意时刻成功规则。

代码和数据可用性。 我们将发布所有训练和评估代码、实验配置、任务集、回合划分，以及 Finger 和 Ball-in-Cup 数据集及其数据收集脚本。

E. 实现与超参数

E.1. 共享训练配置

表7和表 8 列出了共享训练配置（基于 LeWM 骨干网络及其训练计划 [14]），以及环境特定的辅助权重。在模型的时间分辨率下，每个动作 a_t 拼接五个连续的基本动作，因此得到 $d_a = 5d_{\text{prim}}$（在 PushT、Reacher、Finger 和 Ball-in-Cup 上为 10，在 Cube 上为 25），该设置同时用于前向预测器和所有逆向目标。

对于独立 CIA，我们根据 PushT 上 $\{0.01, 0.03\}$ 和 Cube 上$\{0.01, 0.03, 0.1, 0.3, 1, 2, 3, 5\}$ 的验证规划性能，选择 λ 。在 Reacher 上，使用学习得到的条件提议分布进行初步扫描后，$\{0.01, 0.03, 0.1, 0.3, 1\}$ 选出了 $\lambda = 1$；切换到 Reacher 上的固定均匀提议分布时，我们保留 $\lambda = 1$，并将其直接用于 Finger 和 Ball-in-Cup，不再进一步调参。对于 Cube 上的 CIA+CGS，使用单个随机种子在$\lambda_{\text{CIA}} \in \{0.01, 0.03, 0.1\}$ 上进行扫描后，$\lambda_{\text{CGS}} = 0.5$ 选出了 $\lambda_{\text{CIA}} = 0.01$。对于基线方法，我们采用已发表的辅助设置：Fu 和 Ning 的 LeWM-based TS 和 CGS [6], 、Ivashkov 等人的环境特定 SMWM 权重 [10], ，以及 Boylan 和 Hokamp 的固定 AC-MTM 权重和温度 [3]；在 Finger 和 Ball-in-Cup 上，各基线沿用其在 Reacher 上的权重。由于辅助目标的尺度各不相同，不同方法的数值权重无法直接比较。

表 7。**共享　　训练配置**。 各方法采用相同的骨干网络、预测和正则化设置。

编码器	ViT-Tiny，图像块大小 14， 224×224 输入
潜变量	192 维（投影器多层感知机，隐藏层宽度 2048）
预测器	因果变换器，深度 6，16 个注意力头，多层感知机宽度 2048，丢弃率 0.1
历史	3 帧
时间采样	每个模型步 5 个基本动作
SIGReg	权重 $\alpha = 0.09$、1024 个投影、17 个节点
优化器	AdamW [12], 学习率 5×10^{-5}、权重衰减 10^{-3}
学习率调度	在前 1% 的更新次数中进行线性预热，然后余弦衰减至 0
梯度裁剪	范数 1.0
批次大小	128
精度	混合 bfloat16
更新次数	127,960（PushT 上为 139,290）
随机种子	每个环境 5 个

表 8　. **特定环境辅助权重**。 第一行给出原始动作维度。

	Reacher	PushT	Cube	Ball-in-Cup	手指
原始维度 d_{prim}	2	2	5	2	2
TS 权重	0.1	0.01	0.1	0.1	0.1
CGS 权重	0.5	0.01	0.5	0.5	0.5
SMWM（IDM）权重	5	30	1	5	5
AC-MTM（动作-NCE）权重	0.30	0.30	0.30	0.30	0.30
CIA 权重 λ	1	0.01	3	1	1
CIA+CGS： λ_{CIA}	1	0.01	0.01	1	1
CIA+CGS： λ_{CGS}	0.5	0.01	0.5	0.5	0.5

E.2. 基线辅助目标

TS 和 CGS 在每个训练窗口中都作用于全部四个潜变量，并在前 5% 的更新次数内将辅助权重线性递增 [6]。CGS 使用标准化动作；在 Cube 上，它使用夹爪被命令保持闭合时，五个基本动作步长的方块对应的 15 个平移坐标。遵循 Ivashkov 等人 [10], SMWM 根据相邻的潜变量对预测每个动作 $(z_t,\ z_{t+1})$，并对每个窗口中三个潜变量对的逆均方误差取平均值。为使不同方法的防坍缩正则化保持一致，我们在 SMWM 和 AC-MTM 中都加入 SIGReg（权重 0.09）。H 节 还评估了 SMWM-H；它向回归头提供与 CIA 完全相同的因果历史和查询。

对于 AC-MTM， [3], 逆向多层感知器将每对相邻潜变量映射为一个动作查询， u_i 并计算 logits $-\|u_i - a_j\|_2^2/(0.10\, d_a)$，以与批次中的所有归一化动作进行比较（批次大小为 128 时，每个查询对应 384 个候选项）。总损失为 $\mathcal{L}_{\text{pred}} + 0.30\,\mathcal{L}_{\text{action-NCE}} + 0.09\,\mathcal{L}_{\text{SIGReg}}$，用于更新编码器和逆向头，而不会通过直接梯度路径更新前向预测器。由于 AC-MTM 与 CIA 在得分参数化、历史输入、候选数量以及候选分布方面存在差异，第J 节在 CIA 内部隔离候选源效应。（在 Reacher 中，均匀且与历史无关的行为使边缘分布和条件动作分布相同。）

E.3. CIA 实现

得分和所提议的 l　　earned proposal 都以相同的因果 h　　istory 为条件，

$$h_t = \big(z_{t-2}, a_{t-2}, z_{t-1}, a_{t-1}, z_t\big), \tag{54}$$

由三个观测潜变量和两个中间动作组成，下一个潜变量为 $z' = z_{t+1}$。无论 a_t 还是 z' 都不进入 h_t。为防止辅助分支中的 BatchNorm 统计量造成目标帧泄漏，共享投影器使用在当前更新前快照保存的运行统计量，同时保留通过编码器和投影器的梯度。架构细节见表9。

提议分布拟合与更新顺序。在 PushT、Cube、手指和 Ball-in-Cup 上从随机初始化开始，我们首先在冻结世界模型的情况下对 q_η 进行最多 5,000 次更新，并根据验证集负对数似然选择检查点。该提议分布

表 9. **CIA 组件**。提议和评分仅在训练期间使用。

候选项 K	33（已执行动作和 32 个备选项）
行为模型 $q_\eta(a \mid h)$	多层感知机，2 个隐藏层，每层 256 个单元，GELU 动作上的 5分量对角高斯混合分布， 在动作边界处截断
提议优化器	AdamW，学习率 10^{-3}，无权重衰减，梯度裁剪 1.0
评分 $S_\psi(h,\ z',\ a)$	多层感知机，包含 2 个各有 256 个单元的隐藏层，GELU，温度 1
评分优化器	共享世界模型优化器和调度

是一个在动作边界处截尾的 5分量对角高斯混合分布：超出边界的样本会被截断，因此其对数似然由内部密度和边界点质量共同构成。在每个联合训练步骤中，我们从更新前提议分布中采样备选项并将其分离，对q_η 基于分离后的历史特征进行负对数似然更新，然后使用存储的候选项执行世界模型更新。Reacher 使用与其已知行为相匹配的固定均匀提议分布，因此跳过提议分布拟合。

对比更新。 所有 $K = 33$ 候选项均针对同一个观测 z'进行评分，无需进行反事实展开。候选位置与正样本索引会一同置换，而 S_ψ 直接输出 logits，不进行 –对数q_η校正，也不使用外部温度。对比损失的权重恒为 λ，且只更新评分模块和共享编码器／投影器（通过 h_t 和 z'），CIA梯度不会作用于前向预测器和提议模块。潜在预测和 SIGReg 保持不变。规划时仅保留编码器和前向预测器。

E.4. 规划配置

在每个环境中，所有方法都使用相同的终端潜在目标代价和规划器设置（表 10）。尽管前向预测器是在三帧历史上训练的，但每次规划调用都仅从当前观测初始化滚动执行，并使用空的历史动作上下文，再自回归地追加候选动作和预测潜在状态（最多保留最近的三对）。该流程在步骤 0 和 25 时运行。

表10. MPPI 和 CEM 设 定值在每种环境下的方法间共享。最大执行预算为 50 个原 e
始环境步；规划在 第 0 步和第 25 步计算，每个规划跨度为五个模型步（25 个原始步）。

MPPI	128 个样本，30 次迭代，温度为 64（在 PushT 中为 4），初始标准差为 1.0
CEM	128 个样本，30 次迭代，32 个精英样本，初始标准差为 1.0

E.5. 对动作候选数量的敏感性

我们在所有环境中均默认使用 $K = 33$ （1 个已执行动作和 32 个备选项）。为测试候选项数量的敏感性，我们在其他训练和规划设置保持不变的情况下，遍历 $K \in \{9, 17, 33, 65, 129\}$，并在相同的 150 个测试任务上，针对 MPPI 和 CEM 的每种设置各评估五个随机种子。如表 11 所示，将 K 降至 33 以下会损害性能——在 Finger 上尤为明显；而将 K 增至 65 或 129 则没有带来稳定的收益。因此， $K = 33$在规划准确性与评分成本之间实现了良好的平衡。

表11. **对动作候选数量的敏感性** K。MPPI–CEM 平均成功率（%），报告为五个训练随机种子下的均值 $\pm$ 标准误。 $K = 33$列沿用了主实验结果。

	候选项总数 K				
环境	9	17	**33**	65	129
PushT	85.8 ± 0.8	85.7 ± 0.8	86.5 ± 0.6	87.0 ± 0.7	86.6 ± 0.7
Reacher	75.2 ± 0.6	74.7 ± 0.3	74.7 ± 0.7	75.5 ± 0.7	75.3 ± 0.3
OGBench-Cube	68.8 ± 2.5	73.1 ± 0.3	73.6 ± 2.5	70.3 ± 2.9	69.4 ± 3.3
手指	24.9 ± 1.2	28.3 ± 0.5	29.9 ± 0.8	29.6 ± 0.9	29.7 ± 1.6
Ball-in-Cup	36.0 ± 2.2	33.7 ± 2.6	39.3 ± 1.3	36.0 ± 3.2	39.9 ± 0.3

E.6. 诊断探针配置

第 F 节中的结果盲探针使用宽度为 512 的三隐藏层多层感知机，并采用 AdamW 训练（学习率 10^{-3}，无权重衰减，批次大小为 512，使用余弦衰减），最多训练 20,000 次更新，并采用验证交叉熵早停（耐心值为 5,000 次更新）。所有报告的比较均使用完全相同的预算。

F. 无下一潜变量的候选识别

实验方案。为了量化在不观察结果的情况下区分备选项的难易程度 Z'，我们在 Ball-in-Cup 和 Cube 的每个环境中分别冻结三个条件CIA编码器，并针对第 5.5节中从每个条件源抽取的 33 个候选项，训练全新的结果盲判别器（使用冻结模型学习值中的样本q_η，用于条件源）。请注意，这些诊断实验在不同条件源之间共享相同的冻结条件CIA编码器，而第 J节中的规划消融实验则针对每个条件源重新训练独立的世界模型。

每个判别器接收与以下三种上下文之一配对的候选动作：无上下文（$X = \varnothing$）、带有过去动作的物理状态历史（$X = H_{\text{phys}}$），或带有过去动作的模型因果潜变量历史（$X = H$）。没有任何判别器接收 Z'。我们使用 307,200 个训练集、10,240 个验证集和 20,480 个测试集回合不相交窗口，采样具有随机化候选位置的新训练备选项，并依据验证交叉熵选择检查点（第 E.6节）。我们先在每个编码器随机种子内对三个探针初始化取平均，再对三个编码器随机种子取平均（与模型无关的均匀候选专用探针和边缘候选专用探针，则在每个环境中仅进行一次、基于三个初始化取平均）。

指标和结果。我们 测量留出交叉熵增益相对于随机机会的表现（对数 $33 \approx 3.497$ 纳特；随机准确率 $1/33 \approx 3.03\%$），

$$\widehat{\Delta}_X = \log 33 - \widehat{\mathcal{L}}_X. \tag{55}$$

表 12 展示了不同候选来源之间清晰的递进关系。均匀替代项几乎无需任何上下文即可被完美识别（$\geq 99.8\%$ 准确率）。边缘采样消除了这种候选专用捷径（$\sim 3.3\%$ 准确率），但提供历史后又恢复了较高的可识别性（60.0% 在 Ball-in-Cup 上，以及 92.9% 在具有潜在历史的 Cube 上）。历史条件采样（q_η）相较于边缘采样显著降低了潜变量历史交叉熵增益——在 Ball-in-Cup 上从1.82降至1.00纳特，在 Cube 上从 3.32降至1.22纳特。

在 Cube 上，对于条件替代项，潜变量历史探针的评分略低于候选专用探针（35.8% 对比 36.7%）。这反映的是有限容量优化，而不是添加历史导致信息丢失；在一个随机种子上进行更长时间的训练会缩小这一差距，但无法完全消除该差距。

残余提议线索。尽管条件采样抑制了基于历史的捷径，但仍会留下可测量的仅候选项线索：Ball-in-Cup上的准确率为29.6%，Cube上为36.7%（使Cube的仅候选项交叉熵增益从0.03纳特提高到1.27纳特）。因此，学习值q_η 在实践中优于边缘采样（见J节），但尚未实现精确行为匹配。在表 2中，提议分散度和边界频率捕捉的是特定的分布统计量，而非完整的条件分布（有关直接的仅历史 R^2 估计值，请参见K节）；残余不匹配仍会通过行为先验比进入公式（14）$\widehat{}$。在表12的所有条目中，Δ_X 的最大标准误为0.031纳特。正如B.5节所述，每个探针的增益估计的是其自身的性能，而不是对 $b(q)$的上界，因此，从训练得分的增益中减去该增益并不能得到经过认证的互信息界。

与训练得分的关系。在这相同的三种子测试队列上，训练得到的条件 CIA 得分在 Ball-in-Cup 上达到 89.3% 的准确率（交叉熵为 0.285 纳特），在 Cube 上达到 88.8%（0.339 纳特）。将 Z' 替换为另一个测试窗口的下一个潜变量后，准确率分别骤降至 2.8% 和 3.7%。因此，训练得分高度依赖 Z'，尽管新探针仍能在没有它的情况下提取残余线索。请注意，报告的 Ball-in-Cup 准确率略有差异（表 2 中为 89.7%，表16 中为 90.3%，此处为 89.3%），反映了各自不同的评估队列：分别是在验证集上评估三个模型、在验证集上评估五个模型，以及在测试集上评估三个模型。

G. 表征可访问性

我们评估了 LeWM、SMWM、AC-MTM 和 CIA 的冻结 PushT 编码器（每种方法使用三个训练随机种子），以考察学习到的表征中物理状态变量的可访问性。

表12。**不含下一个潜变量的识别**。新判别器的留出准确率和交叉熵。交叉熵和增益以纳特为单位；聚合方式如上所述。

环境	备选项	上下文	准确率（%）	CE	$\widehat{\Delta}_X$
Ball-in-Cup	均匀	仅候选项	99.8	0.006	3.491
		物理历史	99.9	0.003	3.494
		潜在历史	99.9	0.005	3.492
	边缘分布	仅候选项	3.3	3.463	0.033
		物理历史	66.9	1.230	2.267
		潜在历史	60.0	1.676	1.821
	条件分布 q_η	仅候选项	29.6	2.594	0.903
		物理历史	59.3	1.613	1.884
		潜在历史	41.5	2.502	0.995
Cube	均匀	仅候选项	100.0	0.000	3.497
		物理历史	100.0	0.001	3.495
		潜在历史	100.0	0.001	3.496
	边缘分布	仅候选项	3.4	3.471	0.026
		物理历史	93.7	0.163	3.333
		潜在历史	92.9	0.177	3.320
	条件分布 q_η	仅候选项	36.7	2.228	1.269
		物理历史	41.2	2.076	1.421
		潜在历史	35.8	2.275	1.222

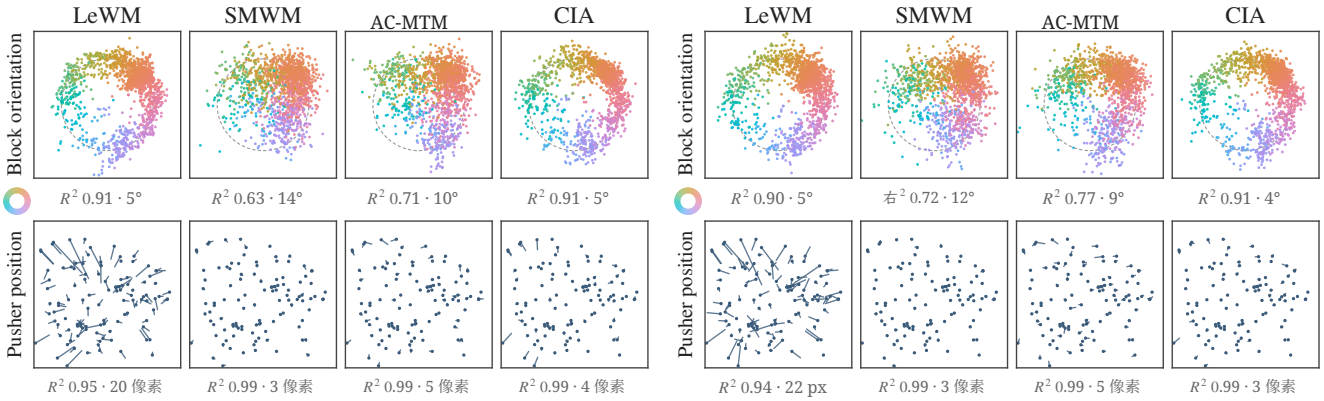


图 6. **另外两个编码器随机种子的读出平面**。 可视化方式与图 1相同。

线性状态读出。 沿用 LeWM 和 SMWM 中的冻结编码器探测方法 [10, 14], ，我们在 100,000 帧训练集上拟合岭回归，在 10,000 帧验证集上调整正则化，并在按回合划分且互不重叠的数据集上进行评估，测试集包含 20,000 帧（使用训练统计量对特征和目标进行标准化）。方块角度编码为 $(\cos\theta,\ \sin\theta)$，并进行聚合 R^2；位置得分则取各坐标得分的平均值 R^2。

表 13 汇总了跨随机种子的测试集 R^2。图 1 在共享的随机测试子集上可视化了每种方法的一个随机种子（上：预测的方块 $(\cos\theta,\ \sin\theta)$；下：真实与预测的推动器位置），图 6 则展示了剩余两个随机种子的结果。三个随机种子的结果高度一致：LeWM 的推动器误差中位数为 20–22 像素，而逆向方法为 2.85–5.48 像素（图中标签四舍五入到最接近的整数）；LeWM 和 CIA 的方块角度误差中位数为 4–5°，SMWM 为 12–14°，AC-MTM 为 7–10°。

解读。 CIA 将精准的推杆位置读出（$R^2 = 0.9905$）与强大的方块朝向读出（$R^2 = 0.918$独特地结合起来，其朝向表现与 LeWM 相当，同时显著优于 SMWM（0.656）和 AC-MTM（0.766）。图 7将这些线性读出与原生潜变量检索诊断并列展示——在未缩放的潜变量平方欧氏距离下，测量最近邻朝向误差（面板 b）和位置与朝向联合一致性（面板 c）。综合来看，

表 13. **PushT 上的线性状态读出**。测试集 R^2，报告为三个冻结编码器种子下的均值 $\pm$ 样本标准差。数值越高越好。

方法	推动器位置	方块位置	方块角度
LeWM	0.9428 ± 0.0061	0.974 ± 0.004	0.910 ± 0.007
SMWM	0.9941 ± 0.0003	0.949 ± 0.007	0.656 ± 0.060
AC-MTM	0.9916 ± 0.0002	0.959 ± 0.005	0.766 ± 0.055
CIA	0.9905 ± 0.0008	0.975 ± 0.003	0.918 ± 0.010

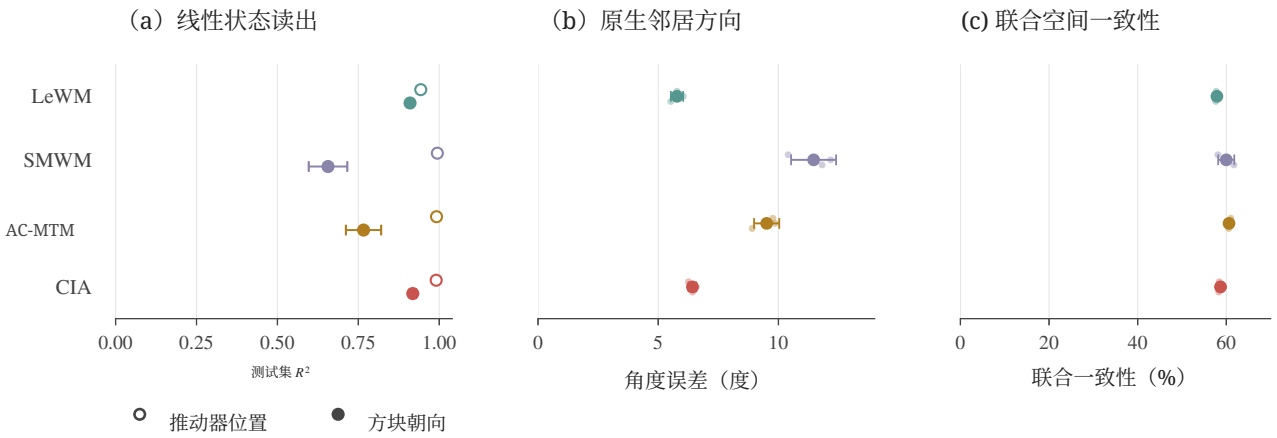


图 7。**PushT 上的表征诊断**。（a）对推动器位置和方块朝向进行线性读出。（b）检索帧的方块角度误差。每个模型都会检索方块池中潜变量与查询潜在表示之间欧几里得距离平方最近的帧。（c）检索帧在推动器位置和方块位置上相距不超过 20 像素、且方块角度误差不超过 $20°$ 的查询所占比例。误差条表示三个冻结编码器随机种子下的样本标准差。联合一致性是检索诊断指标，而非规划成功率。

这些诊断表明，CIA 能够同时保留多个与目标相关的状态分量，即使其原生检索几何本身并非始终表现最佳。

在 PushT 失败分解（第 5.4节）中，这一朝向读出优势伴随着$2.6\%p$：与 AC-MTM 相比，CIA 的平均方块相关失败率高出这一比例。此处，方块相关失败是指以仅方块欧几里得位置误差 $\geq$ 20像素 或环绕方块方向误差 $\geq 20°$结束的失败回合，占所有评估回合的比例。

H. 匹配历史回归逆动力学

为检验 CIA 相较于 SMWM 的优势是否仅仅源于使用了更长的历史，我们训练了 *SMWM-H*，将 SMWM 的成对输入 (z_t, z') 替换为 CIA 所使用的完全相同的因果历史 (h_t, z')。其他所有设置——架构（仅加宽逆向多层感知器的第一层）、损失权重（PushT 上为 30，Cube 上为 1，Reacher、Finger 和 Ball-in-Cup 上为 5）、SIGReg、优化器、数据划分以及五个训练随机种子——均与表 1 （表 7和 8）中的 SMWM 保持一致。匹配 CIA 的历史也使查询保持一致：SMWM-H 仅对每个窗口中的最终转移进行回归（该转移具有完整的三帧历史），而标准 SMWM 则对所有三个相邻状态转移的均方误差取平均值。前向预测器保持不变。

匹配历史和查询并不能缩小与 CIA 之间的差距（表 14）。在全部五个环境中，CIA 仍保持最高的规划器平均成功率，并且总体上超过 SMWM-H $+6.2\%p$。对任务和训练随机种子进行自助法分析（5,000 次百分位数重复抽样）得到该总体增益的 95% 区间为 $[4.0, 8.4]\%p$。

I. Cube上的规划几何诊断

为了理解学习得到的潜变量目标代价如何影响 Cube 上的动作搜索，我们比较表 1 中的冻结 CIA 和 CGS 模型，考察不同规划预算、梯度敏感性和执行模式下的表现。

表14. **使用匹配历史的回归逆动力学**。测试成功率（%），五个训练随机种子的平均值 ± 标准误。平均值是在每个随机种子内合并 MPPI 和 CEM 得出的。总体值为各环境等权重平均值。粗体表示每行中的最高平均值。

环境	规划器	SMWM	SMWM-H	CIA
PushT	MPPI	76.9 ± 1.2	76.9 ± 0.7	$\mathbf{81.9} \pm 1.1$
	CEM	88.3 ± 1.5	89.3 ± 1.1	$\mathbf{91.1} \pm 0.5$
	平均值	82.6 ± 1.2	83.1 ± 0.8	$\mathbf{86.5} \pm 0.6$
Reacher	MPPI	65.2 ± 0.6	65.1 ± 0.5	$\mathbf{67.3} \pm 1.0$
	CEM	80.4 ± 1.2	79.3 ± 1.9	$\mathbf{82.0} \pm 0.8$
	平均值	72.8 ± 0.8	72.2 ± 0.8	$\mathbf{74.7} \pm 0.7$
Cube	MPPI	68.1 ± 0.7	68.0 ± 0.9	$\mathbf{74.7} \pm 2.3$
	CEM	69.2 ± 0.4	67.5 ± 0.9	$\mathbf{72.5} \pm 2.7$
	平均值	68.7 ± 0.3	67.7 ± 0.8	$\mathbf{73.6} \pm 2.5$
手指	MPPI	19.1 ± 0.9	$\mathbf{19.7} \pm 0.5$	16.7 ± 0.9
	CEM	27.7 ± 2.0	30.0 ± 0.9	$\mathbf{43.2} \pm 1.1$
	平均值	23.4 ± 1.1	24.9 ± 0.6	$\mathbf{29.9} \pm 0.8$
Ball-in-Cup	MPPI	15.3 ± 0.8	14.1 ± 1.5	$\mathbf{20.7} \pm 1.6$
	CEM	29.2 ± 1.6	36.4 ± 1.4	$\mathbf{58.0} \pm 2.0$
	平均值	22.3 ± 0.9	25.3 ± 0.8	$\mathbf{39.3} \pm 1.3$
总体平均值		53.9	54.6	**60.8**

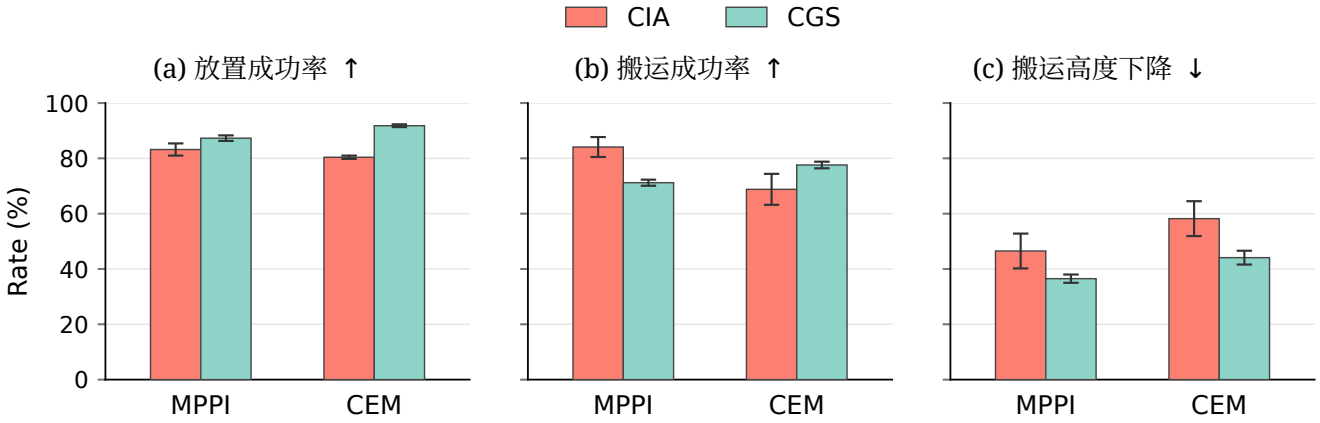


图8。 **在 Cube 上搬运并放置** 。 五个训练随机种子的均值 ±标准误，每个随机种子包含 93 个放置任务和 34 个搬运任务。成功率 s 测量 50 个步骤；高度 下降表示在搬运任务的前 25 个步骤中，Cube 穿过 3 厘米以下的位置。

首先，我们在每个随机种子的 200 个验证任务上，每次仅改变一个规划预算参数。将 MPPI 的样本数量从 32 增加到 512（迭代 30 次），可将 CGS–CIA 的平均成功率差距从 5.4%p 缩小至 1.9%p，且大部分增益在样本数达到 128 时便已实现。然而，在基于梯度的规划（GD）下，CIA 从 66.1%（迭代 10 次）提升至 72.2%（迭代 30 次），随后在 72.7%（迭代 200 次）处趋于平台期；相比之下，CGS 则从 66.3% 稳步提升至 79.1%。因此，CIA 能从更广泛的采样中受益，但延长局部梯度细化带来的收益很小。

规划代价的动作敏感性。为检查哪些动作坐标获得优化信号，我们记录了前 25 步规划中前 100 次迭代的潜在目标代价梯度（一个随机种子，350 个留出任务：200 个验证集任务和 150 个测试集任务）。我们对各次迭代和时间步的归一化动作梯度平方求和，计算每个坐标在每个任务中的占比，并在所有任务上取平均值。在 CIA 中，夹爪和偏航角坐标分别仅占总梯度平方的 1.15% 和 0.04%，而在 CGS 中则分别占 7.16% 和 2.31%。因此，CIA 的规划梯度高度集中于平移，

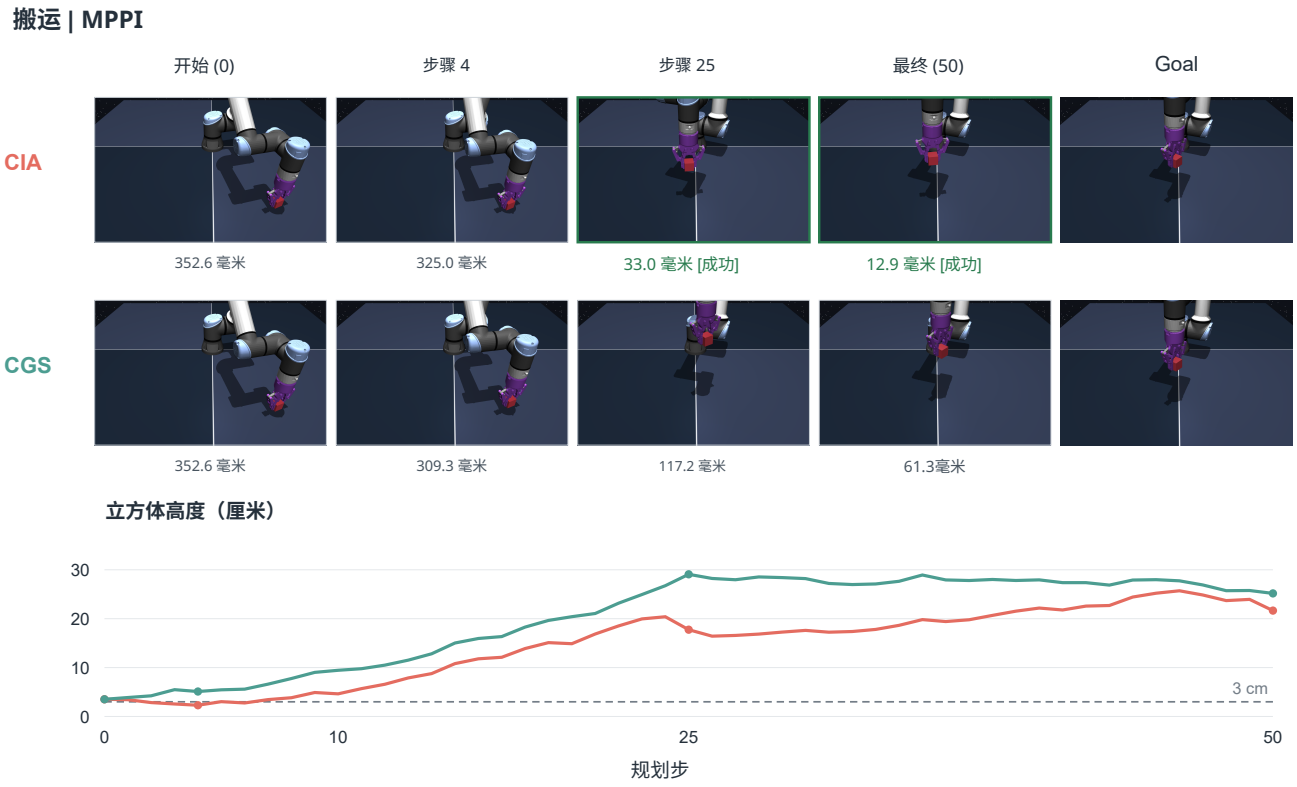


图9。**搬运：高度下降不一定意味着失败**。具有相同起点和目标的匹配的MPPI滚动轨迹。在CIA下，立方体短暂下降至低于3厘米高度阈值，但随后到达目标；在CGS下，立方体保持抬高状态，却未命中目标。数字表示立方体位置误差；绿色边框表示误差在40毫米成功阈值以内。曲线表示立方体高度随时间的变化。

这使得抓取和方向调整获得的相对信号较弱。值得注意的是，CGS 更高的夹爪和偏航敏感性是间接产生的，因为其几何对齐项仅监督平移坐标。

搬运和放置过程中的失败。我们分析每个随机种子下350个留出任务中的MPPI和CEM滚动执行，将高度为 >3厘米且起点—目标位移为 >40 毫米的立方体定义为抬高的立方体。这得到34个搬运任务（起点和目标均抬高）以及93个放置任务（起点抬高，目标低于3厘米）。如图8 所示，在两种规划器下，CGS均取得更高的平均放置成功率；而搬运性能则因规划器而异：在MPPI下CIA领先，在CEM下CGS领先。在搬运任务的前25个规划步中，CIA允许立方体更频繁地下降至低于3厘米，但暂时的高度下降并不总是导致失败（图9）；相反，在放置任务中降低立方体也不能保证到达目标（图10）。

与目标的关系。CIA 通过逆向评分对编码器进行监督，使动作具有可识别性，但规划依赖前向预测器和欧氏潜在目标距离。为动作识别所保留的信息未必会均匀映射到潜变量距离上，也未必能通过多步滚动轨迹顺畅传播，因为 CIA 既不会直接更新预测器，也不会约束规划代价的几何结构。MPPI 差距逐渐缩小、GD 平台期以及平移主导的梯度都指向了这一差异——这促使我们将动作识别（CIA）与几何正则化（CGS）结合起来。

J. 替代动作来源的补充结果

此处报告 Ball-in-Cup、Cube 和 Finger 候选来源消融实验中针对不同规划器的结果（第 5.5 节）。每个 CIA 变体均使用五个训练随机种子和 32 个备选项，其他所有超参数、初始化和数据顺序均保持不变，并在相同的 150 个测试任务上进行评估（时域为 25 步，在 50 步预算内重新规划一次）。LeWM 和 SMWM 被

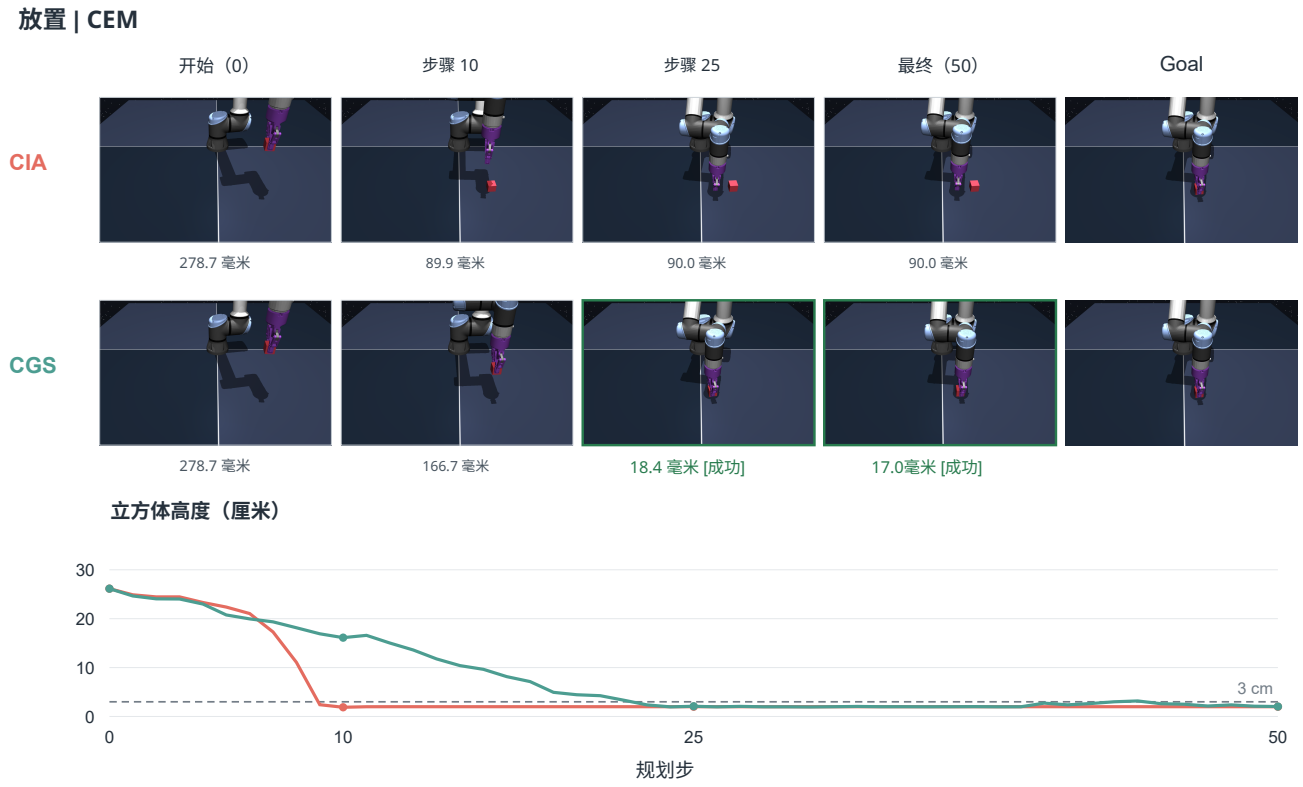


图10。**Place：仅降低立方体是不够的**。具有相同开始状态和目标的匹配 CEM 推演。两种方法都会降低立方体，但只有 CGS 能到达目标位置。数字表示立方体位置误差；绿色边框表示误差在40毫米成功阈值以内。曲线表示立方体高度随时间的变化，虚线表示3厘米。

作为基线方法纳入以供参考。根据第 H 节中的任务与随机种子自助法，条件分布相对于边缘分布的增益在 Ball-in-Cup 上的 95% 区间为 $[1.2, 14.5]$，在 Cube 上为 $[0.1, 12.1]$，在 Finger 上为 $[-1.3, 5.7]$（%p）。

历史条件化替代项在三个环境中均取得最高的规划器平均成功率（表15）。在 Ball-in-Cup 和 Cube 中，条件采样在两个规划器下都优于均匀采样和边缘采样替代项：在全部五个 Ball-in-Cup 随机种子和五个 Cube 随机种子中的四个上均超过边缘采样。在 Finger 中，条件采样同样在规划器平均结果下超过全部五个随机种子的边缘采样；这得益于 CEM 的显著提升（43.2% 对比 36.7%），尽管在 MPPI 下均匀采样领先（20.0%对比 16.7%）。

训练得分对下一个潜变量的依赖性。为了测试各变体的训练得分是否确实依赖于 z'，我们评估了在同一批次的有效查询中随机置换 z' 前后的验证准确率（排除自身匹配，同时固定历史和候选项）。表16 显示，即使 z' 不匹配，均匀备选项得分仍然非常准确（Ball-in-Cup 上为 86.9%，Cube 上为 98.3%），这证实了严重的结果盲区捷径。相比之下，条件得分降至随机水平（2.8% 和 3.9%），而边际得分仅保持略高于随机水平的准确率（5.4% 和 5.9 %）。值得注意的是，条件采样的观测分类准确率最低，却取得了最高的规划成功率——这表明，与在简单负例上取得高准确率相比，区分更难、且符合历史的备选项能提供更有用的表征监督。

K. 基于历史和下一个潜变量的动作可预测性

为了将命题 1 与我们的基准编码器联系起来，我们测量仅凭历史信息预测动作的难易程度、 Z' 在回归和概率目标下带来的增益，以及这些预测增益是否与规划成功相关。我们冻结主实验中的编码器（LeWM、SMWM、AC-MTM 和 CIA），并分别训练回归和概率逆向头，其输入为 H 或 (H, Z')。

表15。**规划器特定的替代源消融**。测试成功率（%），五个随机种子上的平均值 ± 标准误。平均值在每个随机种子内计算。

方法 / 备选项	MPPI	CEM	平均值
Ball-in-Cup（$\lambda = 1$）			
LeWM	16.1 ± 0.6	38.1 ± 1.6	27.1 ± 0.8
SMWM	15.3 ± 0.8	29.2 ± 1.6	22.3 ± 0.9
CIA, 均匀	17.7 ± 1.3	28.9 ± 2.9	23.3 ± 1.6
CIA, 边缘分布	17.9 ± 2.0	45.3 ± 5.4	31.6 ± 2.9
CIA, 条件分布	$\mathbf{20.7} \pm 1.6$	$\mathbf{58.0} \pm 2.0$	$\mathbf{39.3} \pm 1.3$
Cube（$\lambda = 3$）			
LeWM	61.7 ± 0.5	63.6 ± 0.4	62.7 ± 0.4
SMWM	68.1 ± 0.7	69.2 ± 0.4	68.7 ± 0.3
CIA, 均匀	60.0 ± 1.1	61.9 ± 0.7	60.9 ± 0.8
CIA, 边缘分布	68.1 ± 0.9	66.3 ± 0.9	67.2 ± 0.8
CIA, 条件分布	$\mathbf{74.7} \pm 2.3$	$\mathbf{72.5} \pm 2.7$	$\mathbf{73.6} \pm 2.5$
手指（$\lambda = 1$）			
LeWM	14.8 ± 1.0	28.4 ± 0.7	21.6 ± 0.7
SMWM	19.1 ± 0.9	27.7 ± 2.0	23.4 ± 1.1
CIA, 均匀	$\mathbf{20.0} \pm 0.5$	32.7 ± 1.4	26.3 ± 0.5
CIA, 边缘分布	18.9 ± 0.8	36.7 ± 1.1	27.8 ± 0.7
CIA, 条件分布	16.7 ± 0.9	$\mathbf{43.2} \pm 1.1$	$\mathbf{29.9} \pm 0.8$

表16。**下一潜变量替换**。验证识别准确率（%），为每种变体的五个训练模型的平均值。随机机会为3。03%。

备选项	观察到的 z'	已替换 z'
Ball-in-Cup		
均匀	100.0	86.9
边缘分布	93.9	5.4
条件分布 q_η	90.3	2.8
Cube		
均匀	100.0	98.3
边缘分布	98.7	5.9
条件分布 q_η	88.9	3.9

实验方案。所有方法在每个环境中都使用回合不相交集合：204,800 个训练窗口（从 40% 的训练回合中采样）、10,240 个验证窗口和 20,480 个测试窗口；采用 CIA 的三帧因果历史 H。我们评估三个编码器随机种子，每个种子对应两个探针初始化。每个注意力头都是一个包含三层隐藏层的 GELU 多层感知机（宽度为 1024，丢弃率为 0.1），训练 20,000 次更新（批次大小为 1024，使用 AdamW，学习率为 10^{-3} 并采用余弦衰减），输入经过标准化，并使用验证损失检查点保存。概率头采用五分量对角高斯混合分布，并在动作边界处进行截断。我们对每个编码器的两个探针初始化结果取平均值，并报告三个编码器随机种子间的均值和标准误 ± 。

指标。在测试集上，令 V 表示总标准化动作方差，令 M_H 和 M_{HZ} 表示按坐标求和的回归误差。我们报告

$$\widehat{R}^2_H = 1 - \frac{\widehat{M}_H}{\widehat{V}}, \qquad \frac{\widehat{\Delta}}{\widehat{V}} = \frac{\widehat{M}_H - \widehat{M}_{HZ}}{\widehat{V}}, \qquad \widehat{G}_{\mathrm{NLL}} = \widehat{\mathcal{L}}_{\mathrm{NLL},H} - \widehat{\mathcal{L}}_{\mathrm{NLL},HZ}. \tag{56}$$

这里，Δ/V是命题1 中均值偏移项的经验对应量（仅对条件均值敏感），而G_{NLL} （单位为每个动作向量的纳特：$d_a = 25$ 在 Cube 上为10，其他环境中为）是 $I(A;\ Z' \mid H)$的经验对应量，同时也捕捉离散程度、模态结构和边界质量的变化。由于这两个指标处于不同的尺度上，我们在每个指标内部比较编码器，而不跨指标进行比较。

表17。**冻结表征上的动作可预测性**。测试集结果为三个编码器随机种子的均值 ± 标准误，在每个编码器内对两次探针初始化取平均后计算。显示为 0。000 的标准误低于 0。0005。

环境	编码器	$\widehat{R}^2_H$	$\widehat{\Delta}/\widehat{V}$	$\widehat{G}_{\mathrm{NLL}}$ （纳特）
PushT	LeWM	0.840 ± 0.000	0.070 ± 0.000	1.657 ± 0.035
	SMWM	0.830 ± 0.001	0.085 ± 0.001	2.091 ± 0.008
	AC-MTM	0.836 ± 0.000	0.079 ± 0.001	1.777 ± 0.054
	CIA	0.839 ± 0.000	0.068 ± 0.004	1.880 ± 0.034
Reacher	LeWM	-0.002 ± 0.000	0.191 ± 0.000	1.363 ± 0.013
	SMWM	-0.002 ± 0.000	0.192 ± 0.000	1.389 ± 0.014
	AC-MTM	-0.002 ± 0.000	0.192 ± 0.000	1.399 ± 0.018
	CIA	-0.001 ± 0.000	0.193 ± 0.000	1.421 ± 0.011
Cube	LeWM	0.735 ± 0.001	0.123 ± 0.001	3.582 ± 0.049
	SMWM	0.740 ± 0.000	0.188 ± 0.000	5.022 ± 0.027
	AC-MTM	0.746 ± 0.000	0.181 ± 0.000	5.227 ± 0.038
	CIA	0.753 ± 0.000	0.120 ± 0.001	4.375 ± 0.052
手指	LeWM	0.348 ± 0.001	0.219 ± 0.001	1.737 ± 0.013
	SMWM	0.359 ± 0.001	0.232 ± 0.001	1.850 ± 0.009
	AC-MTM	0.359 ± 0.000	0.231 ± 0.001	1.857 ± 0.008
	CIA	0.356 ± 0.001	0.230 ± 0.001	1.871 ± 0.008
Ball-in-Cup	LeWM	0.818 ± 0.000	0.049 ± 0.000	1.939 ± 0.021
	SMWM	0.813 ± 0.001	0.056 ± 0.001	1.748 ± 0.095
	AC-MTM	0.814 ± 0.001	0.056 ± 0.001	1.744 ± 0.029
	CIA	0.810 ± 0.000	0.052 ± 0.001	1.838 ± 0.011

历史的可预测性与下一潜变量的增益。仅凭历史信息，就能解释 PushT、Cube 和 Ball-in-Cup 上 74–84% 的动作方差，在 Finger 上约为 35%，而在 Reacher 上几乎为零（与均匀随机动作一致；表17）。这证实了命题1的实际相关性：在现实的离线数据集中，大部分动作方差在观察 Z'之前就已经可以被预测。即便如此，观察 Z' 仍能提升每个环境和每个编码器上的回归性能与似然——对于 CIA，将仅使用历史信息时剩余的均方误差在 Ball-in-Cup 上降低约 27%，在 PushT 上降低 42%，在 Cube 上降低 49%。

比较不同编码器的回归与似然。 回归增益和似然增益分别反映下一潜变量可预测性的不同方面。在 PushT 和 Cube 上，CIA 在保持归一化回归增益相近的同时，取得了高于 LeWM 的似然增益，而 SMWM 和 AC-MTM 则同时提升了两个指标。在 Ball-in-Cup 上，LeWM 获得了最高的似然增益，而 Finger 和 Reacher 上的差异较小。（由于每个编码器都会同时改变 H 和 Z'，不同编码器之间相匹配的回归增益并不意味着条件均值相同；第 C 节将保持均值不变的信息单独剥离出来。）

与规划的关系。 无论是回归增益还是似然增益，都无法可靠地预测下游规划成功。在 PushT、Cube 和 Ball-in-Cup 上，SMWM 和 AC-MTM 的回归增益均高于 CIA，但 CIA 在 PushT 和 Ball-in-Cup 上胜过二者，并在 Cube 上胜过 SMWM（同时在 Cube 上与 AC-MTM 的表现非常接近；表 1）。类似地，SMWM（PushT、Cube）和 LeWM（Ball-in-Cup）较高的似然增益，并未转化为更高的规划成功率；在 PushT 上，SMWM 和 AC-MTM 较高的回归增益还伴随着方块朝向读出的退化（表 13）。与我们的候选来源消融实验（表 15 和 16）一致，这些发现表明，原始逆向可预测性并不是规划质量的代理指标；应根据辅助监督所保留的、与目标相关的状态区分来评估它。

L. 定性示例

我们展示了从各方法中相同初始状态和目标状态开始的匹配规划轨迹，突出了成功的 CIA 轨迹以及 Cube 上仍存在的一种失败模式。每一帧均依据第D节中的标准进行评分：绿色边框



图11。**PushT：到达推块器和方块目标**。在步数 0、10、25 和 50 时匹配的 CEM 推演，目标位于右侧。CIA、CGS 和 CIA+CGS 满足最终目标。数值表示拼接后的推块器—方块欧几里得位置误差（按 20 像素归一化）与包裹方块角度误差（按 20° 归一化）中的最大值；小于 1 的数值表示成功。

表示目标满足，最终帧上的红色边框表示在第50步失败。根据任意时刻评估规则，即使滚动执行的最终帧结束时超出容差，也可以更早取得成功。

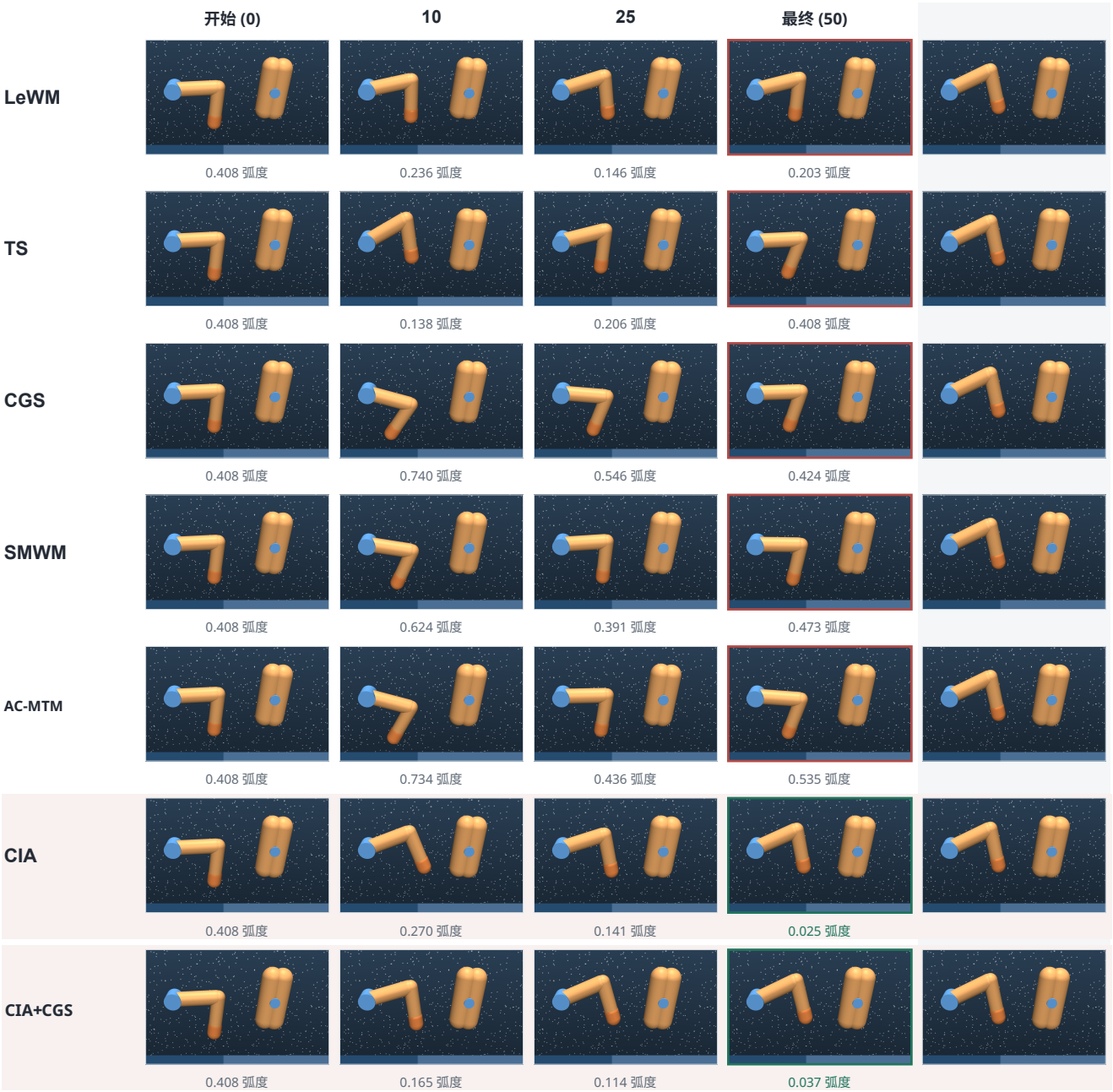


图12。**手指：完成关节构型**。第0、10、25和50步的匹配 CEM 推演，目标位于右侧。CIA 和 CIA+CGS 满足最终目标。数值表示两个手指关节和转盘铰链的最大绝对误差，未进行角度环绕；成功要求误差小于0.05弧度。

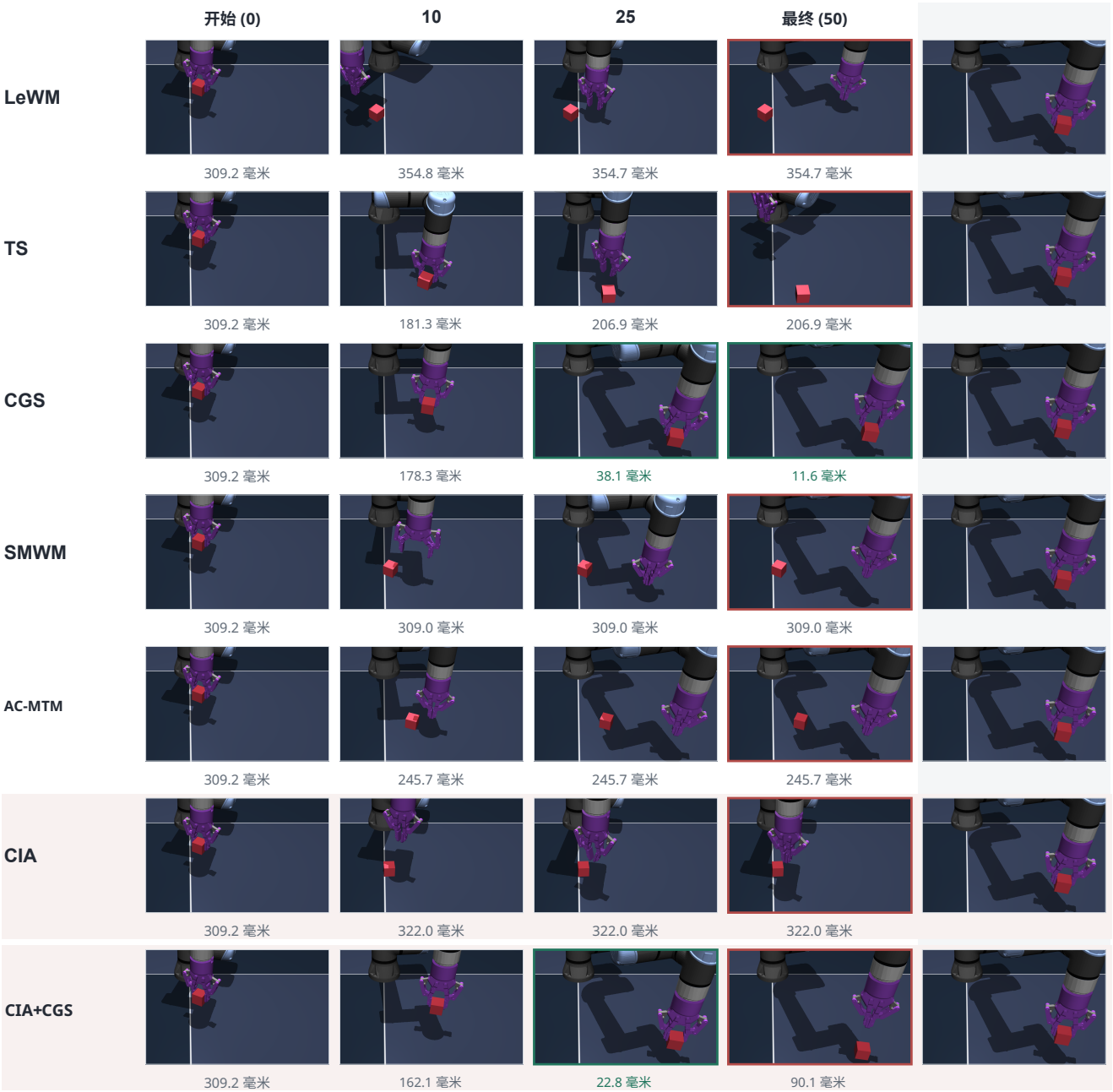


图13。**Cube：物体运输期间的一次剩余失败**。在步骤0、10、25和50处匹配的 CEM 推演，目标位于右侧。CIA 放下了最初抓住的立方体，而 CGS 满足了最终目标。CIA+CGS 在执行期间成功，但最终超出容差。数值表示立方体位置欧氏误差；成功要求误差低于40毫米。

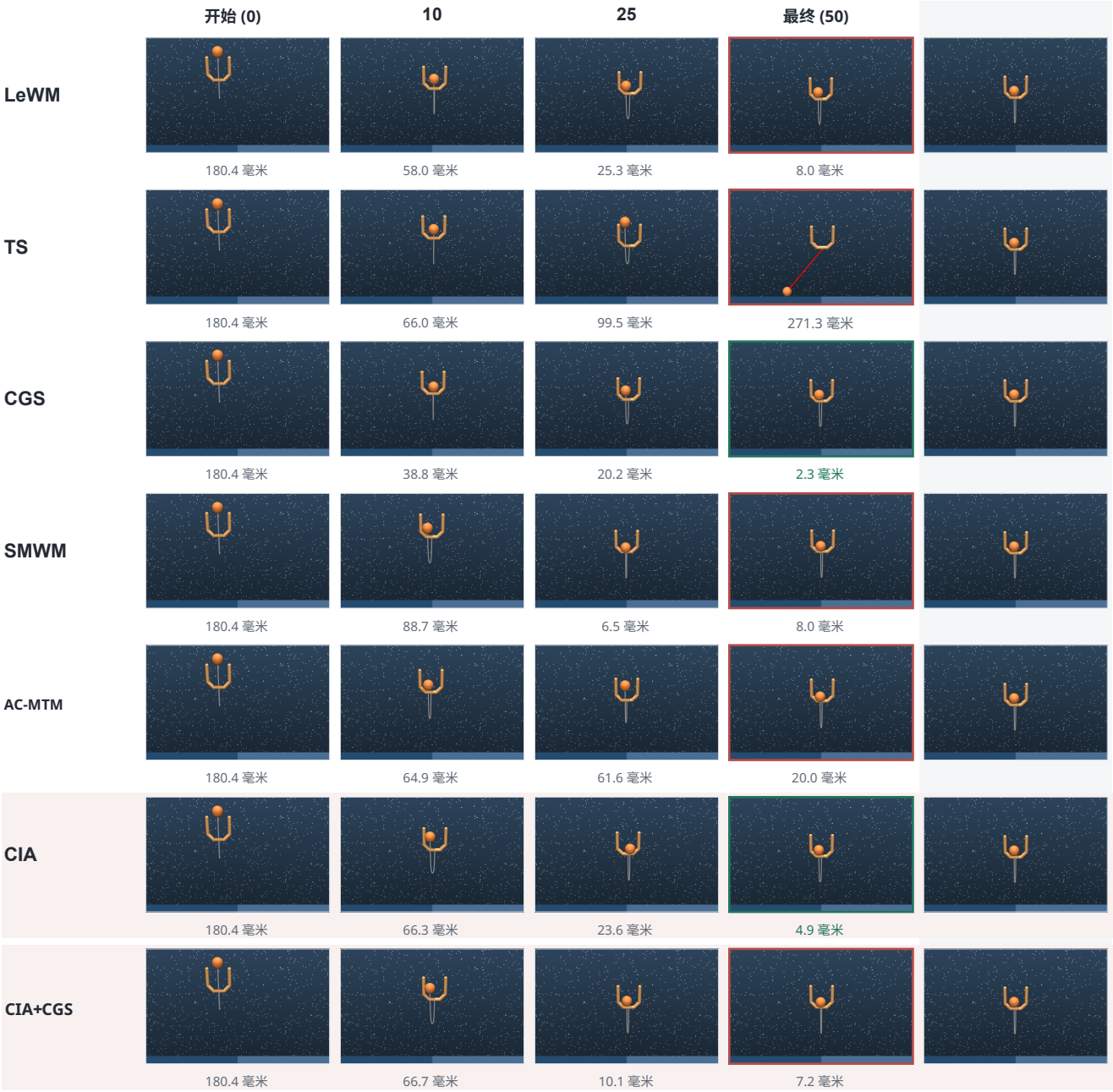


图14。**Ball-in-Cup：从杯外开始到达目标**。第 0、10、25 和 50 步时的匹配的MPPI滚动轨迹，目标位于右侧。CIA 和 CGS 满足最终目标。SMWM、AC-MTM 和 CIA+CGS 更早成功，可能发生在所显示的帧之间，但最终位于容差范围之外。数值表示杯子和球的位置坐标中的最大绝对误差；成功要求每个误差均小于 5 毫米。